HYBRID PHYLOGENIES:

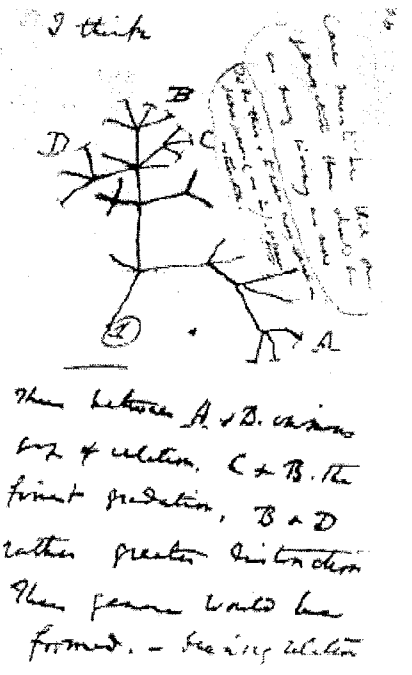
A GRAPH–BASED APPROACH TO REPRESENT RETICULATE EVOLUTION

A thesis
submitted in partial fulfilment
of the requirements for
the Degree of
Doctor of Philosophy in Mathematics
at the
University of Canterbury
by

Mihaela Carmen Baroni

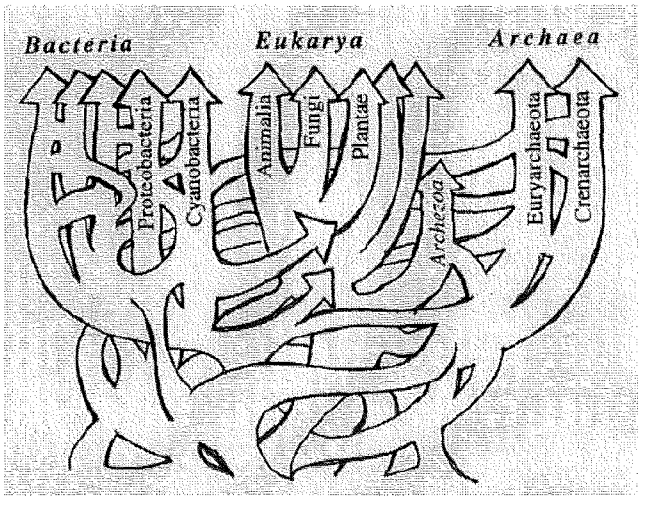
Supervisors: **Professor Mike Steel** and **Dr Charles Semple**

University of Canterbury
Department of Mathematics and Statistics
2004



The first diagram by Charles Darwin
of an evolutionary tree.
(*First notebook on transmutation of species*, 1837)

The tree of life is a twisted, tangled, pulsing entity with roots and branches meeting underground and in midair to form eccentric new fruits and hybrids.
(Lynn Margulis, *The Symbiotic Planet,* Phoenix, London, 1999.)



A reticulated tree, or net, which might
more appropriately represents life's history.
(W. Ford Doolittle, Phylogenetic Classification
and the Universal Tree, *Science*, **284**, 25 June 1999)

Abstract

Although phylogenetic trees provide a useful representation of evolutionary relationships in biology, evolution cannot always be adequately described by the classical tree model. With the increasing recognition of the role of reticulation events (such as hybridization or lateral gene transfer) in evolution, has come the need for developing mathematical models and new tools capable of better representing these phenomena.

In this thesis, we develop a graph–based model for representing reticulate evolution. We define *hybrid phylogenies* as rooted acyclic digraphs with certain properties which attempt to capture the essential biological reality, yet be mathematically tractable. We identify an important subclass—the *regular* hybrid phylogenies (these are isomorphic to the cover digraph of their associated cluster system)—and show that little generality is lost in restricting ourselves to regular hybrids.

This formalism leads to some interesting mathematical problems, with potentially useful applications. One of the main questions is the following: given two rooted binary phylogenetic X–trees $\mathcal{T}$ and $\mathcal{T}'$, how can these trees be *displayed* by a single hybrid phylogeny with a minimum number $h(\mathcal{T}, \mathcal{T}')$ of hybridization events? We relate this number to the *rooted subtree prune and regraft distance* ($d_{rSPR}(\mathcal{T}, \mathcal{T}')$) between $\mathcal{T}$ and $\mathcal{T}'$. A crucial role in studying this problem is played by a particular type of agreement forest, which we call a *good agreement forest*. We show that

$$d_{rSPR}(\mathcal{T}, \mathcal{T}') \leq m_g(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}, \mathcal{T}') \leq |X| - 2,$$

where $m_g(\mathcal{T}, \mathcal{T}') + 1$ denotes the minimum size of a good agreement forest for the two trees.

We describe how the minimum number of hybrid events can be evaluated by reducing the problem to smaller trees. If one tree can be obtained from the other by an appropriate sequence of rSPR operations (corresponding to a maximum good agreement forest), a minimal hybrid can be constructed.

In the last chapter, we introduce and analyse a simple model based on our hybrid setting—the *accumulation phylogenies*. We believe that this model may provide an

alternative technique for reconstructing phylogenetic histories using gene content or other types of genomic markers.

Acknowledgements

First of all I am most grateful to my supervisors Professor Mike Steel and Dr Charles Semple for their continuous advice, support and encouragement.

This thesis was completed with the generous support of a Doctoral Scholarship and of an International Postgraduate Award offered by the University of Canterbury. I would also like to thank the New Zealand Marsden Fund (UOC-005) for conference and travel support.

My warm thanks go to Dr Stefan Grünewald, Dr Katarina Huber and Professor Vincent Moulton for inspiring discussions and comments that have improved my research.

I am very grateful to Professor Solomon Marcus who guided my first steps in research.

Special thanks are due to Professor Cristian Calude for his continuous support and advice.

A warm thank you to Professor Douglas Bridges for encouraging and helping me and my husband to start our PhD studies in New Zealand.

A word of gratitude to Vivien Bridges as well as Neville and Ella Bolsover for their friendship and invaluable help. I would also like to thank my friends Cătălina Iticescu and Liliana Marinescu for their kind and efficient support.

Thanks to the people in the Department of Mathematics and Statistics for providing a friendly and helpful environment.

Finally, I wish to thank my family for their love and patience.

Contents

Chapter 1

Introduction

This thesis develops new mathematical techniques and tools to study reticulate evolution in biology. Although the examples and motivations discussed come from biology, we expect it may provide useful tools for linguistics (for example, for modelling language contact). However, we make no claim to address such questions in this thesis.

1.1 Why reticulate evolution?

Since Darwin's first sketch of an evolutionary tree (1837), biologists have used trees to describe evolutionary relationships between species. However, more recently—and particularly following the analysis of genetic data—it has become increasingly recognized that a network–like pattern is more appropriate to represent phenomena like hybridization in certain plant and fish species and lateral gene transfer in bacteria.

There has been considerable debate in the literature concerning the role of hybridization in evolution. Nearly 20 years ago, Funk wrote [20]

> *It is difficult to overemphasize the importance of hybridization and polyploidy in evolution because they are outstanding features of many plant groups.*

Aspects of detecting and representing hybridization in biology have been discussed by many authors (for example see [13, 34, 41, 44, 45, 47, 49, 50, 57]).

In a recent paper [17], Doolitle stressed the importance of reticulate evolution in the form of lateral gene transfer for the evolution of bacteria. He wrote that

> *Molecular phylogeneticists will have failed to find the "true tree", not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree.*

Reticulate patterns of relationships can also be found in other situations such as the co-evolution of hosts and parasites [43], and historical biogeography [33]. These aspects are discussed in [32].

Legendre [31] summarized the development of biological concepts regarding reticulate evolution as follows:

> *Reticulate patterns of evolution pose a new challenge to evolutionary biologists who have been trained, after Darwin, to believe that the evolution of life could conveniently be summarized and modelled by a branching structure. The existence of reticulated patterns confronts this belief, with two consequences: on the one hand, evolutionary biologists hesitate to study the reticulated facet of evolution because they are reluctant to abandon the paradigm in which they have been trained; on the other hand, those who would like to do so lack an alternative set of tools to represent this new facet of life.*

The increasing need for tools and methods to represent reticulate evolution has given rise to challenging mathematical and computational problems. However, much of the analysis in the biological literature has been somewhat ad–hoc. One such approach was described by Legendre and Makarenkov [33] for inferring a reticulation network ('reticulogram') from an empirical distance matrix. Starting from a tree, one can introduce additional arcs in a heuristic fashion until some stopping criterion

is reached. The method has been applied to examples from biogeography, population microevolution, and hybridization. A similar approach has been taken by Alroy in [3] with a method he calls "continuous track analysis".

Another approach for describing reticulate events has been to apply existing mathematical methods that generate graphs, rather than trees, to biological data. Four such methods—pyramids [16], weak hierarchies [4], splitsgraphs [18] and reticulograms were reviewed by Lapointe (2000) and compared on the same data set. Lapointe [30] concludes his analysis with the remark that "in spite of interesting mathematical properties, the different reticulistic methods will not necessarily produce biologically meaningful results. Model–based techniques should be developed to serve that purpose."

For representing reticulate evolution, particularly lateral gene transfers, Hallett *et al.* [23, 24, 25] have developed a framework for simultaneous identification of duplication and lateral transfer events. A given gene tree T is mapped into a given species tree S by a mapping (called a 'dt-scenario') satisfying certain (biologically motivated) conditions.

While this thesis was being written, new mathematical and computational methods for analysing and representing reticulate evolution were developed [14, 21, 22, 25, 28, 29, 39, 40, 52]. We describe some of these further in Section 2.3. The emphasis in these papers has been mostly algorithmic while this thesis is primarily a mathematical approach.

1.2 Short guide to the thesis

Following this introductory chapter, there are five chapters, a list of references, a list of symbols, and an index for quick referencing.

In Chapter 2, **Preliminaries**, we provide some useful graph–theoretic background, focusing on some concepts from the mathematical foundations of phylogenetics. We also discuss an important tool for understanding and representing retic-

ulate evolution: the rooted subtree prune and regraft operation (rSPR). Section 2.3 is an overview of some mathematical models for representing reticulate evolution.

In Chapter 3, **Hybrid phylogenies**, we introduce our model, a graph–based approach for representing reticulate evolution. We define *hybrids* as rooted acyclic digraphs with certain properties which attempt to capture the biological reality. We identify an important subclass—the *regular* hybrid phylogenies (the hybrids that are isomorphic to the cover digraph of their associated cluster system)—and show that no generality is lost in restricting our model to the regular case. Furthermore, regular hybrids are suitable for a *temporal representation* that more closely reflects the biological reality. We introduce the notion of *display* for hybrids. We show that for any collection of rooted phylogenetic trees, there exists a canonical hybrid, the *cluster union* hybrid, that displays each of the trees in the collection, and we characterize the hybridization vertices of this hybrid. We show that, provided the two trees are sufficiently similar, the cluster union hybrid uniquely determines these trees.

The 'gene trees' for different genes of the same set of species can suggest different evolutionary relationships between species. In such a case, it may be more appropriate to describe the evolutionary picture either by a sequence of trees or by a hybrid phylogeny. In Chapter 4, **Measuring the dissimilarities between trees**, we address an important question of interest for biologists: given two phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on sets of species that faithfully represent the evolution of different parts of species' genomes, how can these trees be displayed by a single hybrid phylogeny with a minimum number of hybridization events? Furthermore, how is this number related to the distance between the two trees measured by the rSPR metric? We introduce three possible measures of hybridization and discuss the relationships between them. We show that for a regular hybrid displaying two trees, the number of hybrid events can be greatly reduced if other species (not sampled in any of the input trees) are permitted. This is biologically motivated since other species (including ones that are extinct now) may have been involved in the evolution in the past. However, we prove that displaying two phylogenetic X–trees by a hybrid with the same set X of leaves is equivalent to displaying trees by a regular hybrid whose set of leaves contains X.

A crucial role in studying this problem is played by the notion of agreement forest. It has been shown in [11] that a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ corresponds to the rSPR distance between the two trees. We introduce a particular type of agreement forest that we call a (maximum) *good agreement forest*, which corresponds to the minimum number of hybrid events required for displaying the trees by a hybrid phylogeny.

Although the difference between the rSPR distance and the hybridization measure can be large for trees with a large number of leaves, we prove that they are equal for small enough trees. Also, we show how a regular hybrid can be constructed in the case where one tree is obtained from the other by a single rSPR operation.

The rSPR distance seems to be a good measure if one is interested in knowing how "far apart" two rooted binary phylogenetic trees are. On the other hand, we believe that our hybridization measure is more appropriate if one is interested in reconstructing a 'minimal' evolutionary history in a consistent way. We address this problem in Chapter 5, **How to construct a minimal hybrid—an example from biology**. We describe how we can evaluate the minimum number of hybrid events by reducing the problem to smaller trees. Given two trees such that one can be obtained from the other by a 'good' sequence of rSPR operations, we show how we can construct a minimal hybrid displaying the two trees. We apply our results to a biological example.

Recently, the gene content of species has been used for reconstructing phylogenies (see [54]) by constructing measures of (dis)similarity based on the amount of genes shared by two species. In the final chapter, **Accumulation phylogenies**, we formalize and analyse a simple mathematical model for the biological situation in which characteristics are passed on to all descendant species. We show that the resulting observed sets of characteristics for the species at the leaves uniquely determine the digraph that describes the evolution of the species, under certain restrictions. Second, we characterize when this digraph is actually a tree.

The results in Chapters 3,4,5 and 6 are new. Chapter 3 is joint work with Mike Steel and Charles Semple. Chapter 6 is joint work with Mike Steel. The part of Chapter 4 regarding agreement forests is joint work with Stefan Grünewald, Vincent

Moulton and Charles Semple. Unless otherwise stated, all other work in this thesis is my own.

Chapter 2

Preliminaries

2.1 Basic concepts and definitions

First, we overview some basic terminology concerning digraphs. For additional background see [4] and [26].

A *directed graph* (or a *digraph*) D is an ordered pair (V, A) consisting of a non-empty set V of *vertices* and a subset A of $V \times V$ of *arcs*. If $a = (u, v)$ is an arc, then u is the *tail* and v is the *head* of a. The arc a is said to be directed from u to v. We denote the set of vertices and arcs of D by $V(D)$ and $A(D)$, respectively. A graph G is the *underlying* graph of a digraph D if G can be obtained from D by replacing each arc with an edge having the same end vertices.

A *directed path* of a digraph $D = (V, A)$ is a sequence

$$p = v_0, a_1, v_1, a_2, v_2, \ldots, v_{k-1}, a_k, v_k$$

of distinct vertices and arcs such that $a_i = (v_{i-1}, v_i)$, for all $i \in \{1, 2, \ldots k-1\}$. If $v_0 = v_k$ then p is a *directed cycle* of D. A digraph is *acyclic* if it has no directed cycles. The *in–degree* (respectively, *out–degree*) of a vertex v of D, denoted $d^-(v)$ (respectively, $d^+(v)$), is the number of arcs of D whose head (respectively, tail) is v.

An acyclic digraph with no underlying parallel edges is *rooted* if there exists a distinguish vertex ρ, called the *root*, such that $d^-(\rho) = 0$ and there exists a directed path from ρ to every vertex of D. Let us observe that, except for ρ, no other vertex has in–degree zero.

Let $D = (V, A)$ be a rooted digraph. For $u, v \in V$, we write $u <_D v$ if there exists a directed path in D from u to v, and $u \leq_D v$ if $u <_D v$ or $u = v$. We say that u is an *ancestor* of v and v is a *descendant* of u. Note that $\leq_D$ is a partial order on the set of vertices of D.

Phylogenetic trees provide a convenient representation of evolutionary relationships in biology. Unless otherwise stated, the terminology and notations in this thesis follow [48].

A *phylogenetic X–tree* (or a *phylogenetic tree on X*) $\mathcal{T}$ is a tree with no degree–two vertices and every *leaf* (vertex of degree one) labelled by a bijective map defined on X. If, in addition, every interior vertex (vertex that is not a leaf) of $\mathcal{T}$ has degree three, then $\mathcal{T}$ is called a *binary phylogenetic tree*. The set X is called the label set of $\mathcal{T}$ and is denoted by $\mathcal{L}(\mathcal{T})$. We can view X as the set of leaves of $\mathcal{T}$ and consequently, denote the leaves of $\mathcal{T}$ by the elements of X.

A *rooted phylogenetic X–tree* is defined in a similar way, except that one interior vertex, which has degree at least two, is distinguished and called the *root*. The remaining interior vertices have degree at least three. A *rooted triple* is a rooted binary phylogenetic tree with label set of size three. The tree consisting of a single–root vertex labelled by the element of a singleton set is also considered a rooted binary phylogenetic tree.

There is a natural bijection between rooted (binary) trees and unrooted (binary) trees. Given a rooted (binary) phylogenetic tree $\mathcal{T}$ and a new leaf $\rho \notin \mathcal{L}(\mathcal{T})$, adjoin ρ to the root of $\mathcal{T}$ and construct the unrooted (binary) tree $\mathcal{T}'$ having the same underlying graph. The operation is reversible. Given an unrooted (binary) tree $\mathcal{T}'$ and $\rho \in \mathcal{L}(\mathcal{T}')$, delete the leaf ρ and its incident edge and root the resulting (binary) tree at the remaining end–vertex of this edge. This correspondence is illustrated in Figure 2.1. The tree $\mathcal{T}_u$ is the unrooted tree corresponding to $\mathcal{T}$.

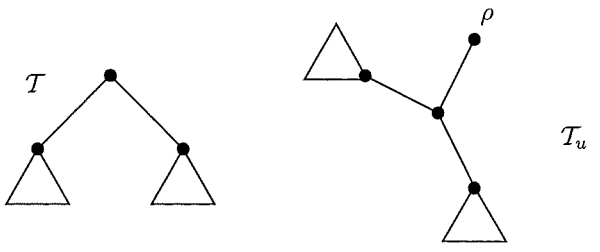


Figure 2.1: A rooted binary phylogenetic tree $\mathcal{T}$ and the corresponding unrooted binary phylogenetic tree.

A rooted tree can be regarded as a rooted digraph by viewing each edge as an arc directed away from the root. Under this interpretation, in a rooted binary tree each vertex has out–degree zero (a leaf) or two (interior vertex).

For any rooted tree $\mathcal{T}$ and u, v vertices of $\mathcal{T}$, we will refer to the unique vertex of $\mathcal{T}$ that is the greatest lower bound of the set $\{u, v\}$ under the partial order $\leq_{\mathcal{T}}$ as the *most recent common ancestor* of u and v in $\mathcal{T}$, and we will denote it by $\mathrm{mrca}_{\mathcal{T}}\{u, v\}$.

In biology, a rooted phylogenetic tree $\mathcal{T}$ on X can describe the evolution of the set X of extant species that label the leaves of $\mathcal{T}$ from a common hypothetical ancestral species at ρ; the interior vertices of $\mathcal{T}$ correspond to further hypothetical ancestral species or to past speciation events, and $\mathrm{mrca}_{\mathcal{T}}\{u, v\}$ is regarded as the most recent shared ancestral species or speciation event.

Let $\mathcal{T}$ be a rooted phylogenetic X–tree and v a vertex of $\mathcal{T}$. Then the set

$$c(v) = \{x \in X : v \leq_{\mathcal{T}} x\}$$

is the *cluster* of v and we denote by

$$c(\mathcal{T}) = \{c(v) : v \in V(\mathcal{T})\}$$

the *set of clusters* of $\mathcal{T}$. The set of clusters $c(\mathcal{T})$ of any rooted phylogenetic X–tree is a *hierarchy* on X, that is, for all $A, B \in c(\mathcal{T})$,

$$A \cap B \in \{\emptyset, A, B\}.$$

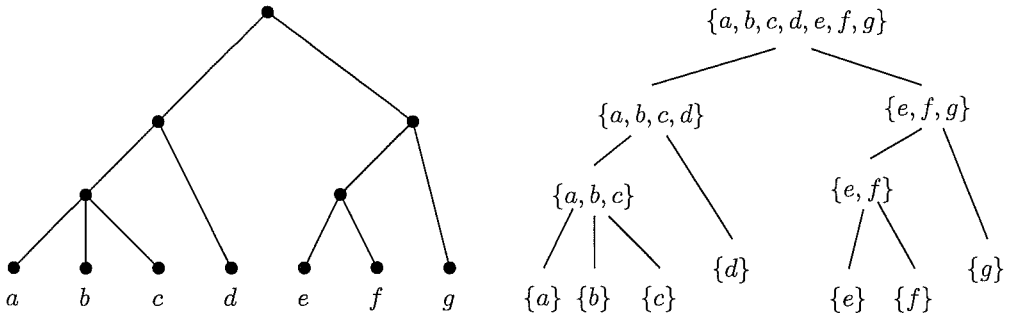


Figure 2.2: A rooted phylogenetic tree and the hierarchy of clusters.

Furthermore, one can easily reconstruct the tree from its set of clusters (see [48]).

Let $\mathcal{T}$ be a phylogenetic X-tree and let X' be a non–empty subset of X. We denote by $\mathcal{T}(X')$ the minimal subtree of $\mathcal{T}$ that connects the vertices of X'. The *restriction* of $\mathcal{T}$ to X', denoted by $\mathcal{T}|X'$, is obtained from $\mathcal{T}(X')$ by suppressing any degree–two vertices. If $\mathcal{T}$ is a binary phylogenetic tree, so is $\mathcal{T}|X'$.

Many results in this thesis are related to an important subtree transfer operation used in phylogenetics: the subtree prune and regraft operation (SPR). For this reason, the next section is entirely dedicated to the SPR operation.

2.2 The subtree prune and regraft operation (SPR)

The subtree prune and regraft operation (SPR) is one of several important tree rearrangement operations used in phylogenetics for the reconstruction and comparison of phylogenetic trees [55]. Furthermore, the subtree prune and regraft operation is useful for modelling the effect of a lateral gene transfer or recombination in genomic data sets (see for example [27, 36, 40, 52]).

The SPR operation in the unrooted case has been considered by Allen and Steel (2001) in [2]:

A *subtree prune and regraft* (SPR) operation on an unrooted binary tree $\mathcal{T}$ is

defined as cutting any edge and thereby pruning a subtree, t, and then regrafting the subtree by the same cut edge to a new vertex obtained by subdividing a pre-existing edge in $\mathcal{T} - t$. A forced contraction[1] is applied in order to maintain the binary property of the resulting tree.

This operation induces a metric on the collection $B(n)$ of binary phylogenetic trees with n leaves. Under this metric, the distance between any two binary phylogenetic trees is defined as the minimum number of operations required to transform one tree to the other. It was shown in [2] that the diameter of this metric space, defined as

$$\text{diam}_{SPR} = \max\{d_{SPR}(\mathcal{T}, \mathcal{T}') : \mathcal{T}, \mathcal{T}' \in B(n)\},$$

is bounded by functions that grow linearly with n. More precisely,

$$n/2 - o(n) \leq \text{diam}_{SPR}B(n) \leq n - 3.$$

The rooted subtree prune and regraft operation (rSPR) can be defined in a similar way, but with the restriction that the pruned subtree should not contain the root of the tree. We describe this operation next.

Let $\mathcal{T}$ be a rooted binary phylogenetic tree. Let (u, a) be an arc of $\mathcal{T}$, and c a vertex of $\mathcal{T}$ in the component containing u. A *rooted subtree prune and regraft* operation (rSPR) $\theta_{a,c}$ on $\mathcal{T}$ is defined as cutting the arc (u, a) and thereby pruning a subtree, t, and then regrafting the subtree by the same cut arc to a new vertex up to c. (See Figures 2.3 and 2.6.) If c is not the root of $\mathcal{T}$, the new vertex is obtained by subdividing the arc ending in c. If c is the root of $\mathcal{T}$, then a new arc (b, c) is added and t is regrafted to b (Figure 2.7). Any resulting degree-two vertices are suppressed in order to maintain the binary property of $\mathcal{T}'$. In the case u is the root of $\mathcal{T}$, the other arc incident with u is contracted.

The *rooted SPR distance* between two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$, $d_{rSPR}(\mathcal{T}, \mathcal{T}')$, is defined as the minimum number of rooted SPR operations required to obtain $\mathcal{T}'$ from $\mathcal{T}$.

[1] A *forced contraction* is an operation on a tree in which a vertex v of degree two is deleted and the two edges incident to v are replaced by a single edge [2].

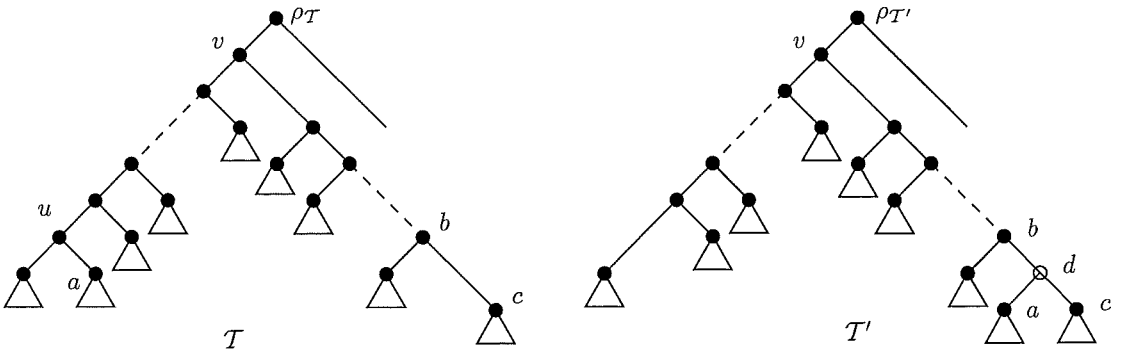


Figure 2.3: An 'across' rSPR operation on a rooted binary tree. The subtree with the root a has been pruned (the arc (u, a) has been cut and u has been suppressed from $\mathcal{T}$) and reattached to the new vertex d. The ancestors of $v = \mathrm{mrca}(a, c)$ and the descendants of u, respectively c are not modified; $c(d) = c(a) \cup c(c)$, $c(u) \cap c(d) \notin \{c(u), c(d)\}$.

It is sometimes unclear in the literature whether regrafting the subtree up to the root of the tree is allowed. We think there are two motivations for allowing this case. First, if this is not part of the definition, then the rSPR operation is not reversible, so it cannot induce a metric on the collection of rooted binary phylogenetic trees. Consider for example the two trees $\mathcal{T}$ and $\mathcal{T}'$ in Figure 2.4. Then $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by a single rSPR operation: the subtree labelled by 4 is pruned and regrafted up to 1. If moving up to the root of the tree is not allowed, at least two rSPR operations are needed to obtain $\mathcal{T}$ from $\mathcal{T}'$: for example, the subtree labelled by 2 is cut and regrafted up to 3, then the subtree labelled by 1 is pruned and reattached up to 2.

Second, observe that the definition we considered is consistent with the definition of the operation in the unrooted case. This can be seen as follows. Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree. Let $\mathcal{T}_u$ be the unrooted binary phylogenetic tree corresponding to $\mathcal{T}$, obtained by adjoining a new vertex labelled by $\rho \notin X$ to the root of $\mathcal{T}$. To each rSPR on $\mathcal{T}$, corresponds an unrooted SPR operation on $\mathcal{T}_u$ with the restriction that ρ is not a vertex of the pruned subtree. On the other hand, a pruned

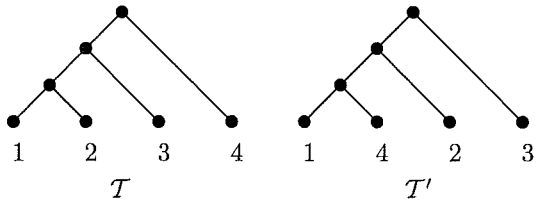


Figure 2.4: The rSPR operation is reversible. To obtain $\mathcal{T}'$ from $\mathcal{T}$ the subtree labelled by 4 is pruned and moved up to 1. By an inverse operation we can obtain $\mathcal{T}$ from $\mathcal{T}'$: the subtree labelled by 4 is cut and reattached up to the root of $\mathcal{T}'$.

subtree of $\mathcal{T}_u$ can be regrafted by subdividing the edge incident with ρ. Note that, as a consequence of the restriction regarding the root, $d_{rSPR}(\mathcal{T}, \mathcal{T}') \neq d_{SPR}(\mathcal{T}_u, \mathcal{T}'_u)$. An example is given in Figure 2.5, where

$$d_{SPR}(\mathcal{T}_u, \mathcal{T}'_u) = 1 < 3 = d_{rSPR}(\mathcal{T}, \mathcal{T}').$$

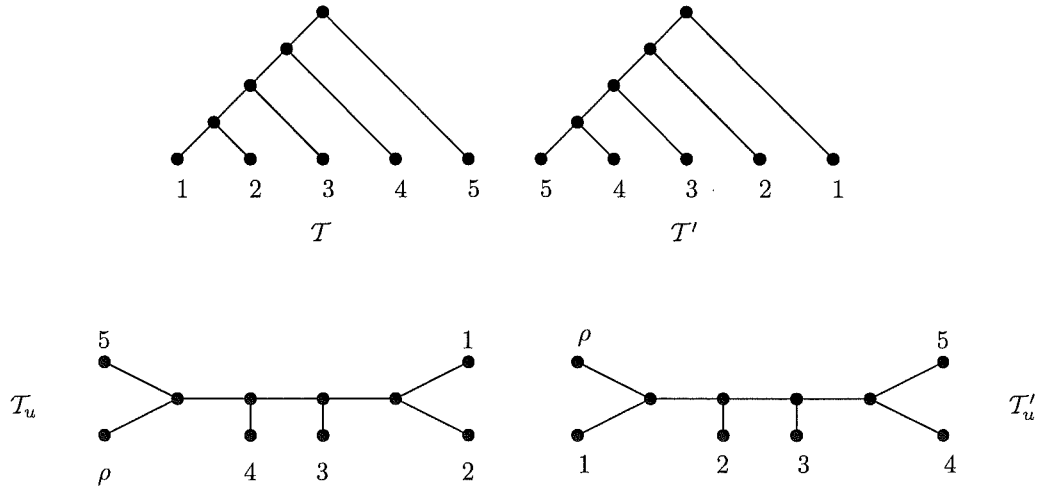


Figure 2.5: Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ with $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 3$ and $d_{SPR}(\mathcal{T}_u, \mathcal{T}'_u) = 1$. The tree $\mathcal{T}'_u$ can be obtained from $\mathcal{T}_u$ by a single SPR operation if the subtree containing ρ is pruned and regrafted.

The rSPR operation can also be described by taking into account the clusters

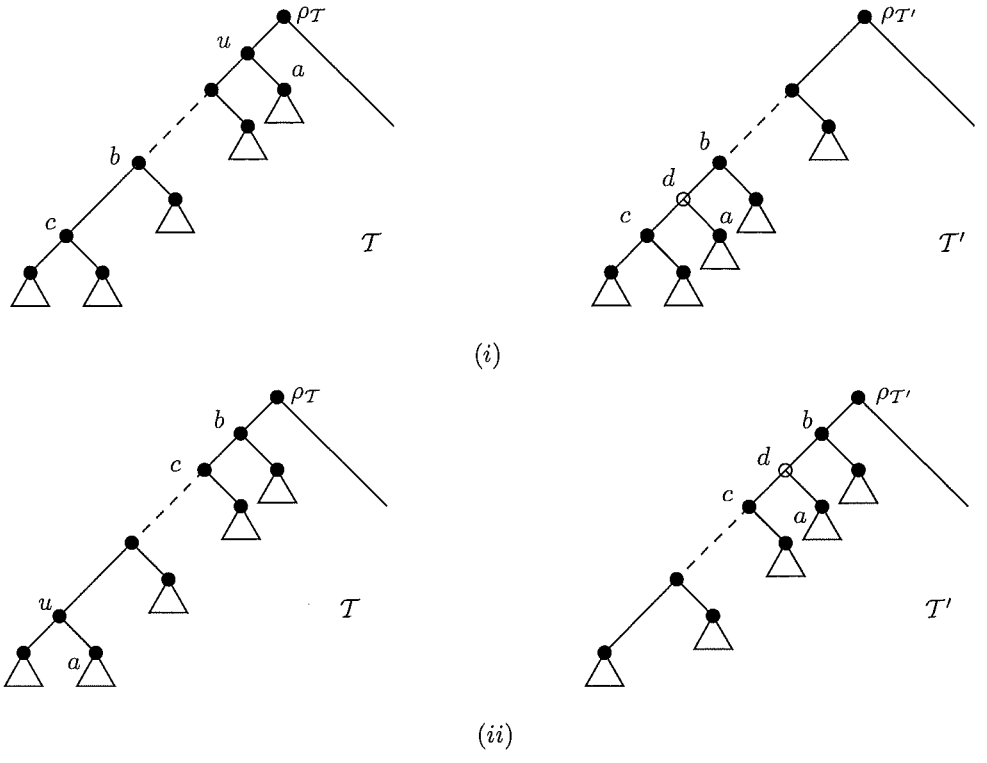


Figure 2.6: (i) A 'down' rSPR operation. The subtree with the root a has been pruned and reattached to the new vertex d, $c(d) \subset c(u)$. (ii) An 'up' rSPR operation; $c \neq \rho_T$. The subtree with the root a has been pruned and reattached to the new vertex d, $c(u) \subset c(d)$.

corresponding to the vertices adjacent to the root of the pruned subtree t in $\mathcal{T} - t$, respectively $\mathcal{T}' - t$. There are three possible cases, described in Figures 2.3 and 2.6.

A similar SPR operation in the rooted case has been considered by Song in [51]. Song proved that, in contrast to the unrooted case [2], the size $|U(\mathcal{T})|$ of the neighbourhood of a phylogenetic rooted binary tree $\mathcal{T}$ depends on the topology of $\mathcal{T}$. Also, the diameter of the metric space of rooted binary phylogenetic trees $RB(n)$ measured using rSPR distance satisfies similar bounds to those for the unrooted case. More precisely, Song proved that

$$n/2 - o(n) \leq \mathrm{diam}_{rSPR} RB(n) \leq n - 2.$$

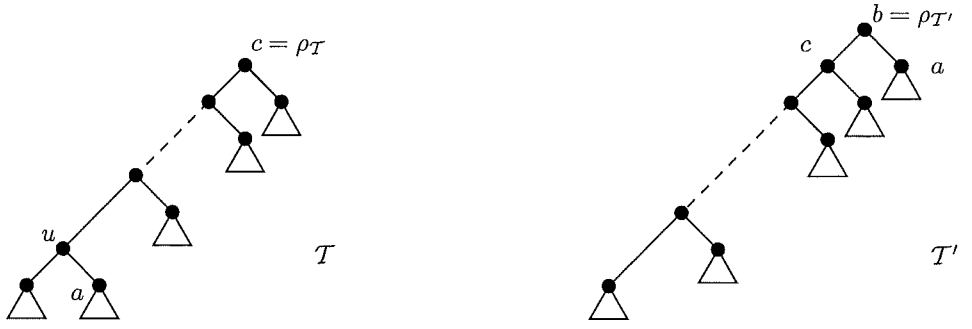


Figure 2.7: An 'up' rSPR operation; $c = \rho_T$. In order to obtain T' from T, the subtree with the root a is regrafted up to the root c of T. The reverse operation, from T' to T, is a 'down' rSPR operation; the subtree with the root a is pruned (the arc (b, a) is cut) and reattached to the component containing b, and the arc (b, c) is deleted.

Recently, Bordewich and Semple (2004) showed that computing rSPR distance is NP–hard [11]. It remains an open problem to determine the complexity of computing the unrooted SPR distance.

2.3 Some other (computational) approaches to represent reticulate evolution

In the last few years, new computational methods have been proposed for representing reticulate evolution, based on directed graphs. We provide a brief overview of these approaches and indicate their relationship to the work presented in this thesis.

In [40], Nakhleh, Warnow and Linder proposed methods for reconstructing accurate evolutionary history in the presence of reticulation events. Their methods are based on the model of "phylogenetic networks". A *binary phylogenetic network* is a directed acyclic graph with exactly one vertex of in–degree zero (the root). The other nodes have: in–degree one and out–degree zero (the leaves), or in–degree one and out–degree two (tree nodes), or in–degree two and out–degree one (reticulation

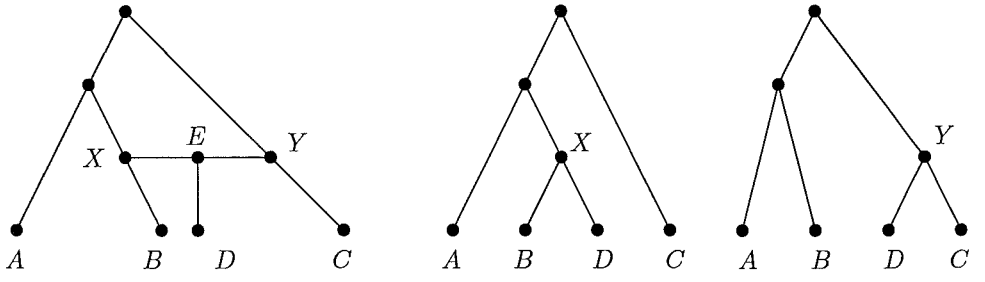


Figure 2.8: A phylogenetic network (the species network) and its two induced (gene) trees, used in [40] to represent hybrid speciation.

nodes). Tree nodes correspond to regular speciation or extinction events, network nodes correspond to reticulation events. The set of edges is partitioned into: tree edges (an edge whose head is a tree node), and network edges (an edge whose head is a reticulation node). Time constraints are imposed on the nodes of the network, such that only nodes that can co–exist in time can be involved in a reticulation event. See Figure 2.8.

Nakhleh *et al.* [40] consider a particular type of phylogenetic network: the "gt-network". A *gt–network* or *galled tree* is a phylogenetic network in which the cycles are node–disjoint. (The reticulation events are considered "evolutionarily independent".) For this special case, Nakleh *et al.* present a polynomial time algorithm for reconstructing a minimal (in terms of the number of reticulation nodes) gt-network that induces two given binary trees.

Galled–trees (see Figure 2.9) were first considered by Wang *et al.* [56]. Gusfield *et al.* [21, 22] formalized and investigated the combinatorial structure of gt–networks and gave an efficient algorithm for the following problem: given a set M of binary sequences, determine if there is a galled–tree that derives M, and if one exists, construct such a network. Galled–trees have also been considered by Jansson and Sung in [29] (there referred to as *level-1 phylogenetic networks*), in the problem of reconstructing such a network from a given set of rooted triples.

In comparison to the above methods, the approach taken in this thesis to represent reticulate evolution is more general as we now describe.

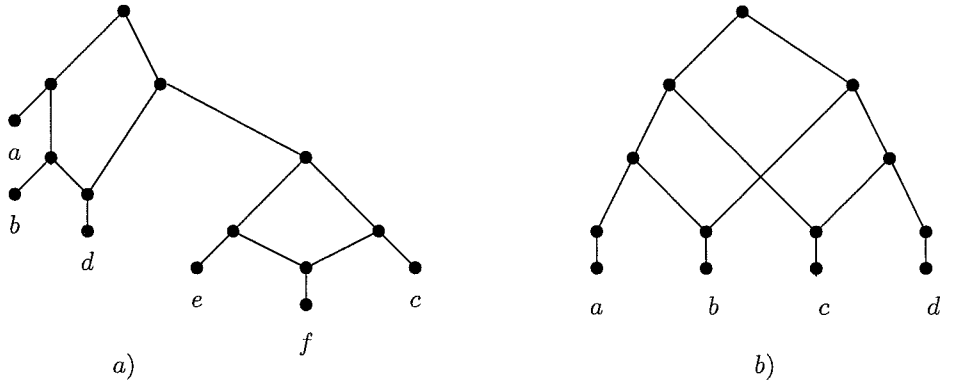


Figure 2.9: a) A galled tree with two 'galls'. b) A network that is not a galled tree.

In our digraph–based approach, we consider *hybrid phylogenies*, a generalization of phylogenetic networks and therefore of the much more restrictive gt–networks. The only restriction imposed on internal vertices of a hybrid phylogeny refers to degree–two vertices: as in the case of rooted phylogenetic trees, no internal vertex of degree two is allowed. We then formalize the notion of what it means for a hybrid phylogeny to *display* a rooted phylogenetic tree. In our setting, the root of the displayed tree can be different from the root of the hybrid, and the leaf set of the tree can be a strict subset of the hybrid leaf set. However, the main difference is that we allow non–independent reticulation events; more precisely, we consider hybrids with non–disjoint cycles and show how such a minimal hybrid that displays two given trees can be constructed.

Chapter 3

Hybrid phylogenies

In this chapter, we introduce our graph–based model for representing reticulate evolution. We formally describe *hybrid phylogenies* as rooted acyclic digraphs satisfying certain constraints and identify an important subclass that will play an important role in our framework—the *regular* hybrids. As we will show later, regular hybrid phylogenies are a natural generalization of rooted phylogenetic trees.

We introduce the notion of *display* for hybrids. We present a graph–theoretic characterization of regular hybrids and use this to prove that, for any hybrid $\mathcal{H}$, there is always a regular hybrid that displays $\mathcal{H}$ and has the same number of hybrid events.

If $\mathcal{H}$ and $\mathcal{H}'$ are two rooted phylogenetic trees, then this definition of display coincides with the usual notion for rooted phylogenetic trees. However, some of the results that hold for trees do not hold in the hybrid setting.

We also consider the particular case of a canonical regular hybrid that displays two rooted phylogenetic trees—the *cluster union* hybrid. We define the *incompatibility graph* for a pair of trees and use this concept to show that the cluster union hybrid uniquely determines the trees in the particular case when the trees are a single rSPR apart.

Related mathematical questions are investigated.

3.1 Basic definitions

In this section we introduce some basic definitions and notations.

Let X be a finite nonempty set. Let $\mathcal{D} = (V, A)$ be a rooted acyclic digraph and ψ a bijective map from X into the set of vertices of V with out-degree 0, such that for all $v \in V - \psi(X)$, $d^+(v) > 1$ whenever $d^-(v) \leq 1$. We say that the ordered pair $\mathcal{H} = (\mathcal{D}, \psi)$ is a **hybrid phylogeny** on X or, simply, a **hybrid**. The set X is called the **label set** of $\mathcal{H}$. The unique vertex of in-degree 0 is called the **root**, the vertices of in-degree at least two are called **hybridization vertices**, and the vertices of out-degree 0 the **leaves** of $\mathcal{H}$. Sometimes we will denote the set $\psi(X)$ of leaves by $\mathcal{L}(\mathcal{H})$. We will often draw the hybrids with their arcs directed downwards, and so omit the arrowheads.

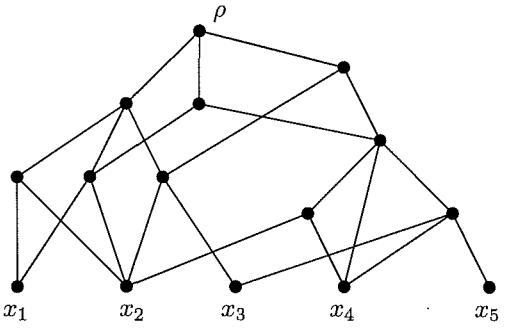


Figure 3.1: A hybrid phylogeny $\mathcal{H}$ with five leaves and seven hybridization vertices. $h(\mathcal{H}) = 10$.

For a hybrid phylogeny $\mathcal{H}$ on X, let

$$h(\mathcal{H}) = \sum_{v \neq \rho} (d^-(v) - 1)$$

be the **hybridization number** of $\mathcal{H}$.

Viewing a hybrid phylogeny as representing the evolutionary history of a collection of present-day species, the hybridization number quantifies the number of associated hybridization events.

Note that $h(\mathcal{H}) \geq 0$, and $h(\mathcal{H}) = |A| - |V| + 1$.[1] We can observe that rooted

[1] This is the cyclomatic number of the underlying graph of $\mathcal{H}$ (see [5]).

phylogenetic trees are special types of hybrid phylogenies. More precisely, a hybrid phylogeny $\mathcal{H}$ is a rooted phylogenetic tree if and only if $h(\mathcal{H}) = 0$.

Two hybrid phylogenies on X, $\mathcal{H} = (\mathcal{D}, \psi)$ and $\mathcal{H}' = (\mathcal{D}', \psi')$, are said to be **isomorphic** if there is a digraph isomorphism $\pi : V \to V'$ with $\psi' = \pi \circ \psi$. We write $\mathcal{H} \cong \mathcal{H}'$ if $\mathcal{H}$ is isomorphic to $\mathcal{H}'$.

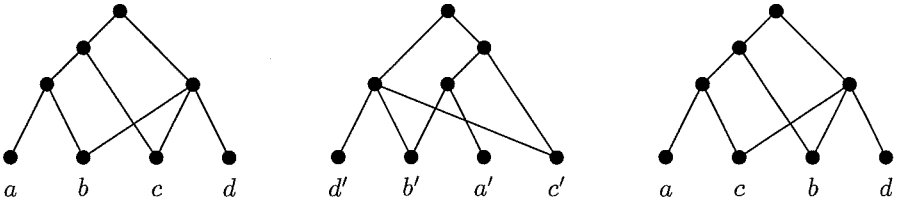


Figure 3.2: The hybrids $\mathcal{H}$ and $\mathcal{H}'$ are isomorphic.

If $\mathcal{H} = (\mathcal{D}, \psi)$ is a hybrid on X and π is an isomorphism between the digraphs $\mathcal{D}$ and $\mathcal{D}'$, then there exists a hybrid phylogeny $\mathcal{H}'$ on X whose corresponding digraph is $\mathcal{D}'$ and such that $\mathcal{H}'$ is isomorphic to $\mathcal{H}$. Furthermore $\mathcal{H}'$ is unique up to a digraph isomorphism. In Figure 3.2, the hybrids $\mathcal{H}$ and $\mathcal{H}'$ are isomorphic, but $\mathcal{H}$ is not isomorphic with $\mathcal{H}''$, although (V, A) and (V, A'') are isomorphic digraphs. However, $\mathcal{H}$ is isomorphic with the hybrid $((V, A''), \phi)$. The maps ψ, ψ', ϕ are defined in the following table.

X	1	2	3	4
ψ	a	b	c	d
ψ'	a'	b'	c'	d'
ϕ	a	c	b	d

3.2 Cluster systems and hybrid phylogenies

Let X be a nonempty finite set. A collection $\mathcal{C}$ of nonempty subsets of X is a **cluster system** on X if $X \in \mathcal{C}$ and $\{x\} \in \mathcal{C}$ for all $x \in X$. We denote by

$$X_{triv} = \{X\} \cup \{\{x\} : x \in X\}$$

the set of **trivial** clusters of X, and we call a cluster C **non–trivial** if $C \notin X_{triv}$.

Let $\mathcal{H} = (\mathcal{D}, \psi)$ be a hybrid phylogeny on X with the vertex set V. Then a cluster system on X can be associated with $\mathcal{H}$ in a canonical way. For $v \in V$, denote by

$$c(v) = \{x \in X : v \leq_{\mathcal{D}} \psi(x)\}$$

the **cluster** corresponding to v. Note that $c(\rho) = X$, $c(v) \neq \emptyset$ for all $v \in V$, $c(\psi(x)) = \{x\}$ for all $x \in X$, and $c(v_1) \subseteq c(v_2)$ whenever $v_2 \leq_{\mathcal{D}} v_1$. The collection

$$c(\mathcal{H}) = \{c(v) : v \in V\}$$

is a cluster system on X, the **set of clusters of** $\mathcal{H}$.

Conversely, given a cluster system $\mathcal{C}$ on X, there is a natural way to obtain a hybrid on X, having $\mathcal{C}$ as its set of clusters. Let $\mathcal{C}$ be a cluster system on X, and consider the Hasse diagram[2] of $\mathcal{C}$ with respect to the partial order relation $\subseteq$; that is, the digraph whose vertex set is $\mathcal{C}$ and for all $C_1, C_2 \in \mathcal{C}$, (C_1, C_2) is an arc precisely if $C_2 \subset C_1$ and there is no $C \in \mathcal{C}$ with $C_2 \subset C \subset C_1$. Define a map $\psi : X \to \mathcal{C}$ by $\psi(x) = \{x\}$ for all $x \in X$, and let $H(\mathcal{C}) = ((\mathcal{C}, A), \psi)$, where A is the set of arcs described above. Then $H(\mathcal{C})$ is a hybrid phylogeny on X, called the **cover hybrid** of $\mathcal{C}$.

Lemma 3.2.1. *Let $\mathcal{C}$ be a cluster system on X and $H(\mathcal{C})$ be the cover hybrid of $\mathcal{C}$.*

(i) If $C, C' \in \mathcal{C}$ then $C' \subset C$ if and only if there is a directed path in $H(\mathcal{C})$ from C to C'.

(ii) $c(H(\mathcal{C})) = \mathcal{C}$.

Proof. (i) Assume that $C' \subset C$ and consider a maximal chain

$$C' = C_1 \subset C_2, \ldots \subset C_n = C$$

where $C_i \in \mathcal{C}$. Then $(C_n, C_{n-1}), \ldots, (C_2, C_1)$ provides the required directed path in $H(\mathcal{C})$ from C to C'. The converse implication follows from the definition of $H(\mathcal{C})$ and the transitivity of inclusion.

[2] This is called the cover digraph in [48].

(ii) We will prove that $c(A) = A$ for all A in $\mathcal{C}$. Clearly, $c(\{x\}) = \{x\}$ for each $x \in X$. Assume now that A has at least two elements. In this case $x \in c(A)$ if and only if there is a path from A to x and, according to part (i), this is equivalent to $\{x\} \subset A$. Therefore the conditions $x \in c(A)$ and $x \in A$ are equivalent. □

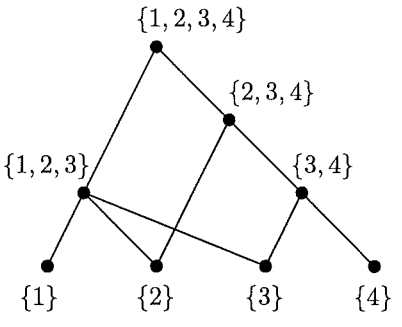


Figure 3.3: The cover hybrid associated to the cluster system $\mathcal{C} = \{\{1\},\{2\},\{3\},\{4\},\{3,4\},\{1,2,3\},\{2,3,4\},\{1,2,3,4\}\}$.

3.3 Regular hybrids

According to Lemma 3.2.1 , given a set of clusters $\mathcal{C}$ of X, the cover hybrid of $\mathcal{C}$ has the cluster set $\mathcal{C}$. As a consequence, $H(\mathcal{C})$ and the cover hybrid of $c(H(\mathcal{C}))$ coincide. In contrast to this result, if we start with an arbitrary hybrid $\mathcal{H}$ and then construct the cover hybrid of $c(\mathcal{H})$, the latter hybrid is not necessarily isomorphic to the initial one. However, we will see that hybrids satisfying this property play a crucial role in our theory.

Let $\mathcal{H}$ be a hybrid phylogeny with the cluster set $\mathcal{C}$. We say that $\mathcal{H}$ is **regular** if the map $v \mapsto c(v)$, from the vertex set of $\mathcal{H}$ to the vertex set of $H(c(\mathcal{H}))$, induces an isomorphism between $\mathcal{H}$ and $H(c(\mathcal{H}))$. For example, every rooted phylogenetic tree is a regular hybrid.

From the definition, it follows that in a regular hybrid phylogeny $\mathcal{H}$ on X, the clusters associated to the vertices are all distinct, and are strictly nested along

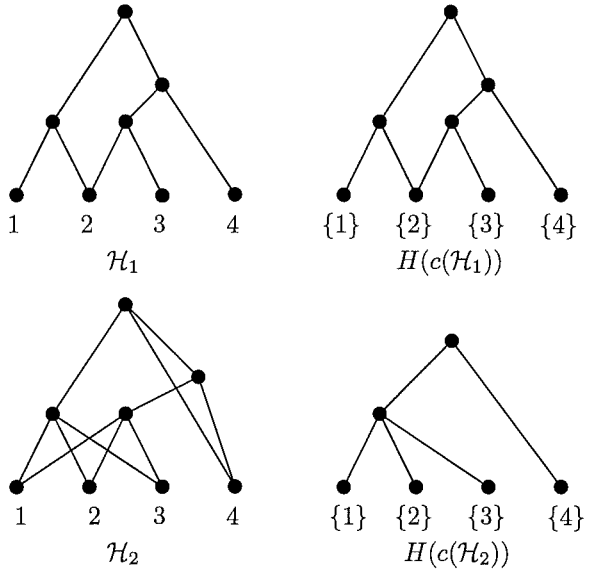


Figure 3.4: A regular hybrid $\mathcal{H}_1$ and a hybrid $\mathcal{H}_2$ that is not regular.

any directed path. Consequently, the longest directed path in any regular hybrid phylogeny on X has at most $|X|$ vertices. Also, a regular hybrid phylogeny has no vertices of out–degree 1.

Two isomorphic hybrids on X have the same cluster set. In general, the converse is not true. Indeed, it is sufficient to consider a non–regular hybrid $\mathcal{H}$. Then $\mathcal{H}$ and $H(c(\mathcal{H}))$ have the same set of clusters but they are not isomorphic. However, two regular hybrids with the same cluster set are isomorphic.

Proposition 3.3.1. *Two regular hybrid phylogenies on X are isomorphic if and only if they have the same set of clusters.*

Proof. Clearly, $c(\mathcal{H}_1) = c(\mathcal{H}_2)$ whenever $\mathcal{H}_1 \cong \mathcal{H}_2$. Conversely, suppose that $\mathcal{H}_1$ and $\mathcal{H}_2$ are two regular hybrid phylogenies with $c(\mathcal{H}_1) = c(\mathcal{H}_2) = \mathcal{C}$. From the definition of regularity, it follows that $\mathcal{H}_1 \cong H(\mathcal{C}) \cong \mathcal{H}_2$. □

The last result enables us to identify two isomorphic regular hybrid phylogenies on X. Furthermore, in this case, we may assume without loss of generality that $X = \mathcal{L}(\mathcal{H})$.

Having identified the isomorphic hybrids on X, let $Reg(X)$ be the family of the regular hybrid phylogenies on X and define the mapping d by

$$d(\mathcal{H}_1, \mathcal{H}_2) = |c(\mathcal{H}_1) \triangle c(\mathcal{H}_2)| \quad (\mathcal{H}_1, \mathcal{H}_2 \in Reg(X)),$$

where $\triangle$ denotes the symmetric difference of two sets.[3] It follows from Proposition 3.3.1 that d is a metric on $Reg(X)$.[4]

The following proposition provides a useful graph–theoretic characterization of regular hybrids.

Proposition 3.3.2. *Let $\mathcal{H}$ be a hybrid phylogeny. Then $\mathcal{H}$ is regular if and only if $\forall v_1, v_2 \in V(\mathcal{H})$, $v_1 \neq v_2$, the following conditions hold.*

(R_1) $c(v_1) \neq c(v_2)$

(R_2) *If* $c(v_2) \subset c(v_1)$, *then* $v_1 <_{\mathcal{H}} v_2$.

(R_3) *If there exist two distinct directed paths connecting v_1 and v_2, neither of them is an arc.*

Proof. If $\mathcal{H}$ is regular then conditions (R_1)–(R_3) are clearly satisfied. Now, suppose that $\mathcal{H}$ is a hybrid phylogeny verifying (R_1)–(R_3). Let $\mathcal{C} = c(\mathcal{H})$. Since $\mathcal{H}$ satisfies (R_1), the map $v \mapsto c(v)$ from $V(\mathcal{H})$ to $\mathcal{C}$ is bijective. We show that this map induces an isomorphism from $\mathcal{H}$ to the cover hybrid $H(\mathcal{C})$.

Let (v_1, v_2) be an arc of $\mathcal{H}$. Then $c(v_2) \subset c(v_1)$ and from the definition of $H(\mathcal{C})$ it follows that there is a directed path in $H(\mathcal{C})$ from $c(v_1)$ to $c(v_2)$. Assuming that this path does not consist of a single arc, then there exists a vertex $c(u)$ in $\mathcal{H}(\mathcal{C})$ such that $c(v_2) \subset c(u) \subset c(v_1)$. The condition ($R_2$) entails the existence of a directed path in $\mathcal{H}$ from v_1 to v_2 that contains u. Then there exist two directed paths from v_1 to v_2, one of which consists of a single arc. This contradicts the fact that $\mathcal{H}$ satisfies condition (R_3). Hence $(c(v_1), c(v_2))$ is an arc of $H(\mathcal{C})$.

[3] The symmetric difference of sets A and B, denoted $A \triangle B$, is the set $(A - B) \cup (B - A)$.

[4] This metric, when restricted to rooted phylogenetic X–trees, is the well-known 'Robinson-Foulds' metric [46].

Now assume that $(c(v_1), c(v_2))$ is an arc of $H(\mathcal{C})$. Then $c(v_2) \subset c(v_1)$ and, as $\mathcal{H}$ satisfies (R_2), there is a directed path in $\mathcal{H}$ from v_1 to v_2. This path must consist of a single arc. Assuming the contrary, it follows that there exists a vertex u on this path, distinct from the vertices v_1 and v_2. Then $c(v_2) \subset c(u) \subset c(v_1)$, therefore $(c(v_1), c(v_2))$ is not an arc of $\mathcal{H}$, a contradiction. □

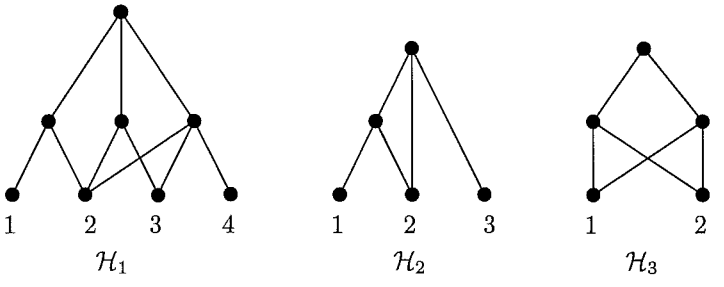


Figure 3.5: Three hybrids to prove that the conditions (R_1)–(R_3) in Proposition 3.3.2 are independent.

Note that the conditions (R_1)–(R_3) in Proposition 3.3.2 are independent. To prove this let us consider the three hybrids in Figure 3.5. The hybrid $\mathcal{H}_1$ satisfies (R_1) and (R_3) but not (R_2). The hybrid $\mathcal{H}_2$ satisfies (R_1) and (R_2) but not (R_3). The hybrid $\mathcal{H}_3$ satisfies (R_2) and (R_3) but not (R_1).

Given a hybrid $\mathcal{H}$, let us consider now the map c from $V(\mathcal{H})$ to $c(\mathcal{H})$ that assigns to each vertex v its corresponding cluster $c(v)$. Clearly, c is one–to–one if and only if it satisfies condition (R_1) in Proposition 3.3.2. Moreover, if we consider $V(\mathcal{H})$ and $c(\mathcal{H})$ as partially ordered sets with respect to the relations $\leq_{\mathcal{H}}$, (respectively, $\subseteq$), condition (R_1) shows that c is a strictly increasing function. If, in addition, $\mathcal{H}$ satisfies (R_2), then c is an order isomorphism between $V(\mathcal{H})$ and $c(\mathcal{H})$.

A hybrid $\mathcal{H}$ that satisfies the conditions (R_1) and (R_2) will be called **almost regular**. We denote by $\mathcal{A}(X)$ the family of the almost regular hybrids on X. Given an almost regular hybrid, we can obtain a regular hybrid in a canonical way.

Proposition 3.3.3. *Let $\mathcal{H} = (\mathcal{D}, \varphi)$ be an almost regular hybrid phylogeny on X and A the set consisting of all arcs (u, v) such that there is a path p from u to v with*

$p \neq (u, v)$. If $\mathcal{D}'$ is the digraph obtained from $\mathcal{D}$ by deleting all the arcs of A then $\mathcal{H}' = (\mathcal{D}', \varphi)$ is a regular hybrid on X.

Proof. Let (u, v) be an arc of A. Clearly, the digraph obtained from $\mathcal{D}$ by deleting (u, v) is rooted and acyclic. We prove that it also satisfies the condition $d^+(y) > 1$ for all non-leaf vertices with $d^-(y) \leq 1$. It suffices to examine the vertices u and v. Since $(u, v) \in A$ it follows that there exists a vertex w, adjacent to u, that lies in a path from u to v. The hybrid $\mathcal{H}$ is almost regular, so $c(w) \subset c(u)$ and, therefore there exists another arc (u, z) with $z \notin \{v, w\}$. Consequently, $\mathcal{H}'$ is a hybrid phylogeny on X.

It is straightforward to observe that conditions (R_1) and (R_2) are preserved and $\mathcal{H}'$ also satisfies the condition (R_3). □

The remainder of this section will be dedicated to some combinatorial results. First, let us calculate the number of non–isomorphic hybrid phylogenies on a given set X. Clearly, if $|X| = 1$, there is only one way to construct a hybrid on X: the trivial graph whose root and single leaf coincide. When X has at least two elements, there are infinitely many non–isomorphic hybrids on X. Furthermore, for each nonnegative integer k we can find at least one hybrid phylogeny on X, having exactly k hybrid events. To prove this, for each $n \geq 2$ and $k \geq 0$, let us consider the rooted digraph $\mathcal{D} = (V, A)$ (Figure 3.6), where $V = \{x_1, x_2, \ldots, x_n, v_0, v_1, \ldots, v_k\}$, and $A = \{(v_k, x_i) : 1 \leq i \leq n\} \cup \{(v_{j-1}, v_j) : 1 \leq j \leq k\} \cup \{(v_j, x_1) : 0 \leq j \leq k-1\}$.

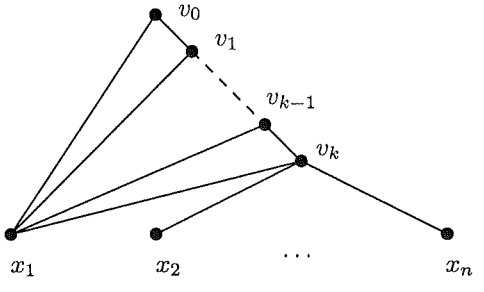


Figure 3.6: A hybrid phylogeny with n leaves and k hybrid events.

If we are restricted to the regular case, Proposition 3.3.1 ensures that the number

of hybrids on X that are mutually non–isomorphic equals the number of cluster systems on X.

Corollary 3.3.4. *For each integer $n \geq 2$, there are exactly 2^{2^n-n-2} non-isomorphic regular hybrids with n leaves.*

Proof. Each cluster system on X can be obtained by adding to X_{triv} any subset, possibly empty, of $2^X - (X_{triv} \cup \{\emptyset\})$. If $|X| = n \geq 2$, then there are 2^{2^n-n-2} elements in $2^X - (X_{triv} \cup \{\emptyset\})$. Therefore this set has exactly 2^{2^n-n-2} subsets. □

The following lemma shows that any hybrid phylogeny on X with a small hybridization number cannot be too large.

Lemma 3.3.5. *Let $\mathcal{H} = ((V, A), \psi)$ be a hybrid phylogeny on X. Then*

$$h(\mathcal{H}) \geq \frac{1}{2}(|V| + 1) - |X|.$$

If in addition $\mathcal{H}$ is regular, then we have:

$$h(\mathcal{H}) \geq |V| - 2|X| + 1.$$

Proof. Let $V_1 = \{v \in \psi(X) : d^-(v) = 1\}$, $V_2 = \{v \in V - \psi(X) : d^-(v) = 1\}$, $V_3 = \{v \in \psi(X) : d^-(v) > 1\}$, and $V_4 = \{v \in V - \psi(X) : d^-(v) > 1\}$. Then $d^+(v) = 0$ for any $v \in V_1 \cup V_3$, and $d^+(v) \geq 2$ for any $v \in V_2$. It follows that

$$|A| = \sum_{v \in V} d^+(v) \geq 2 + 2|V_2| + |V_4|$$

and

$$|A| = \sum_{v \in V} d^-(v) \geq |V_1| + |V_2| + 2|V_3| + 2|V_4|.$$

Adding these two inequalities, and noting that $|V_1| + |V_2| + |V_3| + |V_4| = |V| - 1$ and $|V_1| + |V_3| = |X|$, we obtain $|A| \geq \frac{1}{2}(3|V| - 1 - 2|X| + |V_3|)$, and so, since $h(\mathcal{H}) = |A| - |V| + 1$ and $|V_3| \geq 0$ we obtain

$$h(\mathcal{H}) \geq 1 + \frac{3}{2}(|V| - 1) - |X| - |V| + 1 = \frac{1}{2}(|V| + 1) - |X|.$$

If $\mathcal{H}$ is regular, then $d^+(v) \geq 2$, for each $v \in V - \psi(X)$. It follows that

$$|A| = \sum_{v \in V} d^+(v) \geq 2(|V| - |X|)$$

and

$$h(\mathcal{H}) = |A| - |V| + 1 \geq 2|V| - 2|X| - |V| + 1 = |V| - 2|X| + 1.$$

□

As we have already seen, the number of hybrid events that occur at a vertex (that is, the in–degree of that vertex) could be indefinitely large. However, in the regular case there is an upper bound of this number. We will prove this by using a well–known result in combinatorics, the Sperner's Theorem (see [15]).

A family $\mathcal{F}$ of sets is called a *Sperner family* if no member of $\mathcal{F}$ properly contains any other, that is,

$$A, B \in \mathcal{F} \Rightarrow A \not\subset B \text{ and } B \not\subset A.$$

Sperner's Theorem. *Let $\mathcal{F}$ be a Sperner family of subsets of the n-element set X. Then*

$$|\mathcal{F}| \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

Moreover, the equality holds if and only if $\mathcal{F}$ consists either of all subsets of X of size $\lfloor \frac{n}{2} \rfloor$, or all subsets of size $\lceil \frac{n}{2} \rceil$ (these are the same if n is even).

Proposition 3.3.6. *If $\mathcal{H}$ is a regular hybrid on X with $|X| = n$, then each vertex v of $\mathcal{H}$ satisfies the condition:*

$$|c(v)| = k \Rightarrow d^-(v) \leq \binom{n-k}{\lfloor \frac{n-k}{2} \rfloor}. \tag{3.1}$$

Proof. Let $\mathcal{H}$ be a regular hybrid phylogeny with n leaves. Let v be a vertex of $\mathcal{H}$. We may assume without loss of generality that $X = \{1, 2, \ldots, n\}$ and $c(v) = \{1, 2, \ldots, k\}$. If $d^-(v) = 0$ we have nothing to prove. Suppose now that $d^-(v) = p \geq 1$ and let $u_1, u_2, \ldots, u_p$ be the direct ancestors of v in $\mathcal{H}$. Since $\mathcal{H}$ is regular, for each $i \in \{1, 2, \ldots, p\}$ there exists S_i such that

$$\emptyset \neq S_i \subseteq \{k+1, k+2, \ldots, n\},$$

and

$$c(u_i) = c(v) \cup S_i.$$

Furthermore, the collection

$$\mathcal{F} = \{S_i : i \in \{1, 2, \ldots, p\}\}$$

is a Sperner family of subsets of the $(n-k)$–element set $\{k+1, k+2, \ldots, n\}$. The conclusion now follows from Sperner's Theorem. □

If $\mathcal{H}$ and v are as in Proposition 3.3.6, let us denote by

$$\mathcal{C}_1 = \{C \subseteq 2^X : c(v) \subseteq C,\ |C| = \lfloor \frac{n-k}{2} \rfloor\},$$

and by

$$\mathcal{C}_2 = \{C \subseteq 2^X : c(v) \subseteq C,\ |C| = \lceil \frac{n-k}{2} \rceil\}.$$

Then it is easily seen that $d^-(v) = \binom{n-k}{\lfloor \frac{n-k}{2} \rfloor}$ if and only if the set

$$\{c(u) : u \text{ is a direct ancestor of } v\}$$

equals either $\mathcal{C}_1$ or $\mathcal{C}_2$.

As a consequence of Proposition 3.3.6 , given the number of leaves of $\mathcal{H}$, we can obtain an upper bound of $h(\mathcal{H})$. On the other hand, we can find regular hybrids with n leaves and more than 2^n hybrid events.

Proposition 3.3.7. *Let $X = \{1, 2, \ldots, n\}$ and let $\mathcal{H}_X$ be the regular hybrid with the cluster set $\mathcal{C} = \mathcal{P}(X) - \{\emptyset\}$. Then*

$$h(\mathcal{H}_X) = (2^{n-1} - 1)(n-2).$$

Proof. First, let us observe that the number of vertices of $\mathcal{H}_X$ is $|V(\mathcal{H}_X)| = 2^n - 1$. Then, each vertex v with $|c(v)| = k \geq 2$ has k outgoing arcs. Since there are $\binom{n}{k}$ such vertices, it follows that the number of arcs of $\mathcal{H}_X$ is

$$|A(\mathcal{H}_X)| = \sum_{k=2}^{n} k \binom{n}{k} = n(2^{n-1} - 1).$$

Therefore

$$h(\mathcal{H}_X) = |A(\mathcal{H}_X)| - |V(\mathcal{H}_X)| + 1 = (2^{n-1} - 1)(n-2).$$

□

Corollary 3.3.8. *For $n \geq 7$ there exists a regular hybrid $\mathcal{H}$ with n leaves such that*

$$h(\mathcal{H}) \geq (2^{n-1} - 1)(n-2).$$

Proof. Let $X = \{1, 2, \ldots, n\}$ and let $k \geq 2$. Let $\mathcal{H}$ be the regular hybrid with the cluster set $\mathcal{C} = \mathcal{P}(X) - \{\emptyset\} - \{A \subset X : |A| = k\}$. This is equivalent to considering the regular hybrid $\mathcal{H}_X$ as in Proposition 3.3.7, deleting the vertices corresponding to the clusters of cardinality k and the incident arcs, then, in order to maintain regularity, adding new arcs from vertices corresponding to clusters of size $k+1$ to the vertices whose associated clusters have size $k-1$. It is easily seen that the number of deleted arcs is $n\binom{n}{k}$ and the number of added arcs is $\binom{n}{k-1}\binom{n-(k-1)}{2}$. It follows that

$$h(\mathcal{H}) = h(\mathcal{H}_X) + \binom{n}{k} - n\binom{n}{k} + \binom{n}{k-1}\binom{n-(k-1)}{2}.$$

Denote by

$$\varepsilon_{n,k} = (1-n)\binom{n}{k} + \binom{n}{k-1}\binom{n-(k-1)}{2}.$$

Then note that $h(\mathcal{H}) = h(\mathcal{H}_X) + \varepsilon_{n,k}$ and $\varepsilon_{2p,p} > 0$ if $p \geq 4$ and $\varepsilon_{2p+1,p} \geq 0$ if $p \geq 3$. The conclusion now follows. □

3.4 Displaying hybrids

For phylogenetic trees, the mathematical notion of 'display' captures the concept of preserving ancestral relationships between species and it is fundamental in phylogenetics.

Let X be a subset of X'. A rooted phylogenetic X'–tree $\mathcal{T}'$ *displays* a rooted phylogenetic X–tree $\mathcal{T}$ if $\mathcal{T}'|X$ is a *refinement* of $\mathcal{T}$, that is $\mathcal{T}$ can be obtained from $\mathcal{T}'|X$ by contracting internal edges. If both $\mathcal{T}$ and $\mathcal{T}'$ are rooted binary phylogenetic

trees then the above condition is equivalent to $\mathcal{T}'|X = \mathcal{T}$. The notions of refinement and displaying can be extended in a natural way from trees to hybrids.

Let $\mathcal{H}$ and $\mathcal{H}'$ be two hybrid phylogenies on the same leaf set. We say that $\mathcal{H}$ is a **refinement** of $\mathcal{H}'$ if $\mathcal{H}'$ can be obtained from $\mathcal{H}$ by contracting internal arcs. Let $\mathcal{H}$ be a hybrid on X and $\mathcal{H}'$ be a hybrid on X'. We say that $\mathcal{H}$ **displays** $\mathcal{H}'$ if $X' \subseteq X$ and a rooted subdigraph of $\mathcal{H}$ is a refinement of $\mathcal{H}'$.

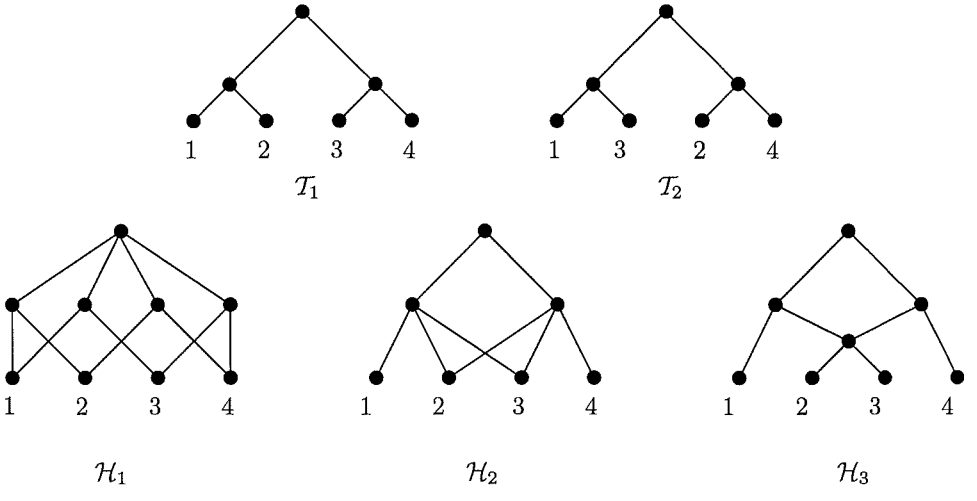


Figure 3.7: Regular hybrids

If $\mathcal{H}'$ is a rooted phylogenetic X'–tree, then $\mathcal{H}$ displays $\mathcal{H}'$ if and only if by deleting certain arcs in $\mathcal{H}$ and suppressing the obtained isolated and out–degree one vertices, a refinement of the given tree is obtained. For example, in Figure 3.7, $\mathcal{H}_1$ and $\mathcal{H}_2$ display both $\mathcal{T}_1$ and $\mathcal{T}_2$. The hybrid $\mathcal{H}_3$ does not display either $\mathcal{T}_1$ or $\mathcal{T}_2$.

If $\mathcal{H}$ and $\mathcal{H}'$ are rooted phylogenetic trees, then the definition of display coincides with the usual definition of display for phylogenetic trees. However, some of the results that hold for rooted phylogenetic trees are not true in the case of hybrids.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X–trees. Then $\mathcal{T}'$ displays $\mathcal{T}$ if and only if $c(\mathcal{T}) \subseteq c(\mathcal{T}')$. The analogous result for two regular hybrids does not hold. For example, in Fig.3.7, $\mathcal{H}_2$ displays $\mathcal{T}_1$ but $c(\mathcal{T}_1)$ is not a subset of $c(\mathcal{H}_2)$. In the same figure, $c(\mathcal{H}_2) \subseteq c(\mathcal{H}_3)$, but $\mathcal{H}_3$ does not display $\mathcal{H}_2$ ($\mathcal{H}_3$ is not a refinement of $\mathcal{H}_2$).

Proposition 3.4.1. *Let $\mathcal{T}$ be a rooted phylogenetic X-tree and let $\mathcal{H}$ be a regular hybrid on X' that displays $\mathcal{T}$. Then there exists a map $\varphi : c(\mathcal{T}) \to c(\mathcal{H})$ with the following properties:*

(i) For each element x of X, $\varphi(\{x\}) = \{x\}$.

(ii) For all $A, B \in C(\mathcal{T})$ with $A \subset B$, $\varphi(A) \subset \varphi(B)$.

(iii) Each cluster A of $\mathcal{T}$ satisfies the condition $A \subseteq \varphi(A)$. In particular, if $X = X'$ then $\varphi(X) = X$.

(iv) Suppose that $X' = X$. If A and $X - A \in c(\mathcal{T})$, then neither $\varphi(A) \subseteq \varphi(X - A)$ nor $\varphi(X - A) \subseteq \varphi(A)$.

Proof. If $\mathcal{H}$ displays $\mathcal{T}$, there exists a rooted subdigraph $\mathcal{T}'$ of $\mathcal{H}$ that is a refinement of $\mathcal{T}$. Let $\varphi_1 : V(\mathcal{T}) \to V(\mathcal{T}')$ be the one–to–one map defined by $\varphi_1(v) = v'$ such that $c(v) = c(v')$ and $d^+(v') \geq 2$ if v is a non-leaf vertex and $d(v') = 1$ if v is a leaf vertex. Then φ_1 is well–defined. Since $\mathcal{T}'$ is a subdigraph of $\mathcal{H}$, each vertex of $\mathcal{T}'$ is a vertex of $\mathcal{H}$. Now let $\varphi : c(\mathcal{T}) \to c(\mathcal{H})$ be the map defined by setting $\varphi(A)$ to be the cluster of $\mathcal{H}$ whose associated vertex is the vertex of $\mathcal{T}'$ that is assigned the vertex of $\mathcal{T}$ corresponding to A under φ_1.

By construction, φ verifies (i). Furthermore, if A and B are clusters of $\mathcal{T}$, and $A \subset B$, since $\mathcal{T}'$ is a rooted subdigraph of $\mathcal{H}$, the vertex corresponding to $\varphi(A)$ is a descendant of the vertex corresponding to $\varphi(B)$. As $\mathcal{H}$ is regular, this implies that $\varphi(A) \subset \varphi(B)$, therefore φ satisfies (ii).

Part (iii) follows from (i) and (ii). Indeed, let $A \in c(\mathcal{T})$; then for all $x \in A$, $\{x\} \subset A$ hence $\{x\} = \varphi(\{x\}) \subset \varphi(A)$. Therefore $A \subseteq \varphi(A)$.

Now, part (iv) is a consequence of (iii). Suppose that $X' = X$ and $A, X - A \in c(\mathcal{T})$ such that $\varphi(A) \subseteq \varphi(X - A)$. Then

$$X = A \cup (X - A) \subseteq \varphi(A) \cup \varphi(X - A) = \varphi(X - A) \subseteq X$$

so $\varphi(X - A) = X$, a contradiction. □

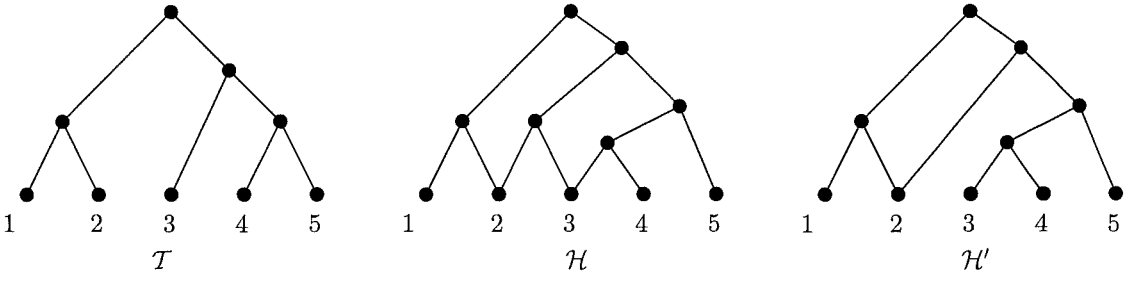


Figure 3.8: The tree $\mathcal{T}$ is displayed by $\mathcal{H}$ but not by $\mathcal{H}'$.

Note that conditions (i) and (ii) are not sufficient for ensuring that a hybrid displays a tree. For the tree $\mathcal{T}$ and the hybrid $\mathcal{H}'$ drawn in Figure 3.8, define $\varphi : c(\mathcal{T}) \to c(\mathcal{H}')$ by

$$\begin{aligned} \{1,2\} &\to \{1,2\} \\ \{4,5\} &\to \{3,4,5\} \\ \{3,4,5\} &\to \{2,3,4,5\} \end{aligned}$$

Then the map φ has the properties (i)–(ii) but $\mathcal{H}'$ does not display $\mathcal{T}$.

Lemma 3.4.2. *Let $\mathcal{T}$ be a rooted phylogenetic tree and let $\mathcal{H}$ be a hybrid that displays $\mathcal{T}$. Then $\mathcal{H}$ displays $\mathcal{T}|U$, for all subsets U of $\mathcal{L}(\mathcal{T})$.*

Proof. Since $\mathcal{H}$ displays $\mathcal{T}$ there exists a rooted subdigraph $\mathcal{T}'$ of $\mathcal{H}$ that is a refinement of $\mathcal{T}$. The minimal rooted subtree of $\mathcal{T}'$ that connects the elements in U is a rooted subdigraph of $\mathcal{H}$ and also, it is a refinement of $\mathcal{T}|U$. It follows that $\mathcal{H}$ displays $\mathcal{T}|U$. □

The converse of Lemma 3.4.2 is not true as the following proposition shows.

Proposition 3.4.3. *Let $X = \{1, 2, \ldots, n\}$, $n \geq 3$. There exist a rooted binary phylogenetic X–tree $\mathcal{T}$ and a regular hybrid phylogeny $\mathcal{H}$ on X such that $\mathcal{H}$ displays $\mathcal{T}|U$, $\forall U \subset X$ with $|U| = k$, $2 \leq k \leq n-1$, but $\mathcal{H}$ does not display $\mathcal{T}$.*

Proof. Let $\mathcal{T}$ be the rooted caterpillar with n leaves. Let $\mathcal{H}$ be the regular hybrid phylogeny with the set of clusters $C(\mathcal{H}) = \{X\} \cup \{C \subset X : |C| \leq n-2\}$. □

Note that for the example in Proposition 3.4.3 we cannot define an order morphism $\varphi : c(\mathcal{T}) \to c(\mathcal{H})$.

Rooted phylogenetic trees are defined by their rooted triples (see [48]). For regular hybrids this is not the case, as the example in Figure 3.9 shows.

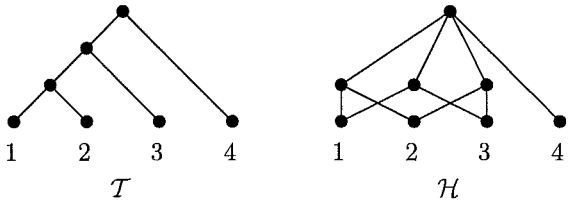


Figure 3.9: The hybrid $\mathcal{H}$ displays every rooted triple of $\mathcal{T}$ but does not display $\mathcal{T}$.

Let us examine now the property: $\mathcal{H}$ displays $H(c(\mathcal{H}))$. Although it is much less restrictive than the regularity condition $\mathcal{H} \cong H(c(\mathcal{H}))$, it is not satisfied in general. A counterexample is given in Figure 3.10.

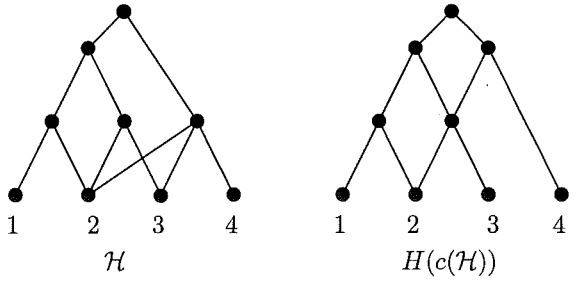


Figure 3.10: A hybrid $\mathcal{H}$ that does not display $H(c(\mathcal{H}))$.

We say that the hybrid $\mathcal{H}$ is **semi–regular** if all the vertices v_1 and v_2 satisfy the condition (R_2):

$$c(v_2) \subset c(v_1) \Rightarrow v_1 <_{\mathcal{H}} v_2.$$

Denote by $\mathcal{S}(X)$ the collection of all semi–regular hybrids on X.

Proposition 3.4.4. *If $\mathcal{H}$ is semi–regular, then $\mathcal{H}$ displays $H(c(\mathcal{H}))$.*

Proof. Let (A, B) be an arc of $H(c(\mathcal{H}))$. Then $B \subset A$. It follows that there exist u and v in $V(\mathcal{H})$ such that $c(u) = A$ and $c(v) = B$. The semi–regularity of $\mathcal{H}$ ensures that there exists a path in $\mathcal{H}$ from u to v. If this path contains another vertex w of $\mathcal{H}$, then either $c(w) = c(u)$ or $c(w) = c(v)$. Let $p = u_0, u_1, \ldots, u_k, v_0$ be a path in $\mathcal{H}$ such that $c(u_0) = A$ and u_0 is minimal with this property, $c(v_0) = B$, and $c(u_i) = A$ for all $i \in \{1, \ldots, k\}$. Consider now the subdigraph of $\mathcal{H}$ induced by the vertices of the paths p, corresponding to the arcs of $H(c(\mathcal{H}))$. It follows that this digraph is a rooted subdigraph of $\mathcal{H}$ and is also a refinement of $H(c(\mathcal{H}))$. Then $\mathcal{H}$ displays $H(c(\mathcal{H}))$. □

The example in Figure 3.11 shows us that semi–regularity is not a necessary condition for $\mathcal{H}$ to display $H(c(\mathcal{H}))$.

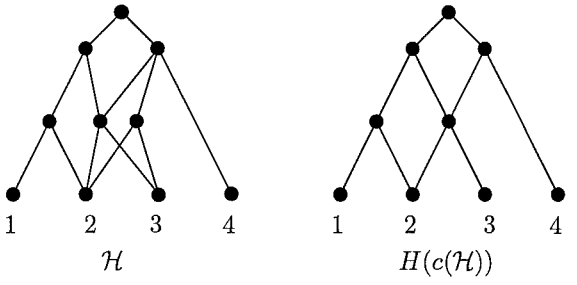


Figure 3.11: The hybrid $\mathcal{H}$ displays $H(c(\mathcal{H}))$ but does not have the property (R_2).

To end this section, let us observe that we have the following hierarchy, with all of the inclusions being strict:

$$\mathcal{T}(X) \subset Reg(X) \subset \mathcal{A}(X) \subset \mathcal{S}(X) \subset \{\mathcal{H} : \mathcal{H} \text{ displays } H(c(\mathcal{H}))\} \subset \mathcal{H}(X).$$

3.5 From non-regular to regular hybrids

Although regular hybrids are a special type of hybrid phylogenies, we show that for any hybrid $\mathcal{H}$, there exists a regular hybrid that displays $\mathcal{H}$ and has the same hybridization number. We describe now how such a hybrid can be obtained.

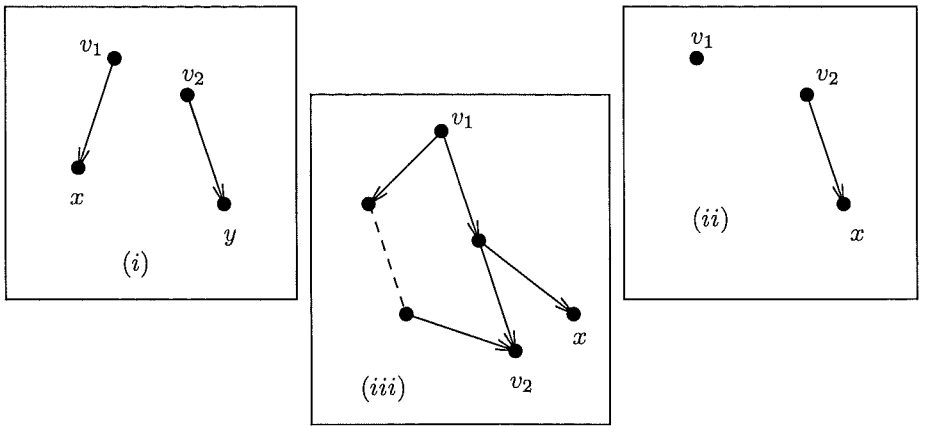


Figure 3.12: Performing operations (i)-(iii).

Let $\mathcal{H}$ be a hybrid phylogeny. For all distinct vertices v_1 and v_2 of $\mathcal{H}$, consider the following sequence of operations.

(i) If $c(v_1) = c(v_2)$, then, for each $i \in \{1, 2\}$, adjoin a new vertex to v_i via a new arc, and assign the new leaf vertex a new label.

(ii) If $c(v_2) \subset c(v_1)$ but there is no directed path from v_1 to v_2, then adjoin a new leaf vertex to v_2 via a new arc and assign the new leaf vertex a new label.

(iii) If there exist two distinct directed paths from v_1 to v_2, one of which is an arc, then subdivide this arc with a single vertex and adjoin a new leaf vertex to the subdividing vertex via a new arc, and assign the new leaf vertex a new label.

Proposition 3.5.1. *Let $\mathcal{H}$ be a hybrid phylogeny and $\mathcal{H}'$ a hybrid obtained from $\mathcal{H}$ by applying operations (i)–(iii) above. Then*

(a) $\mathcal{H}'$ is regular.

(b) $h(\mathcal{H}) = h(\mathcal{H}')$.

(c) Any hybrid displayed by $\mathcal{H}$ is also displayed by $\mathcal{H}'$.

Proof. (a) To prove that $\mathcal{H}'$ is regular, we will show that the conditions (R_1)-(R_3) in Proposition 3.3.2 are satisfied. As each new leaf is assigned a new label, it follows

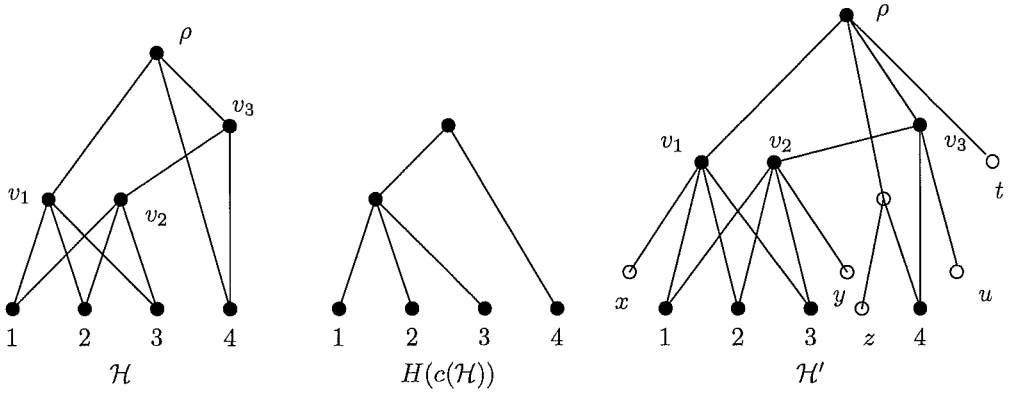


Figure 3.13: A non-regular hybrid phylogeny $\mathcal{H}$ and a regular hybrid obtained from $\mathcal{H}$ by performing operations (i)-(iii).

that for all distinct $v_1, v_2 \in V(\mathcal{H}')$, $c(v_1) \neq c(v_2)$. Because of (ii) in the construction, there always exists a directed path from v_1 to v_2 whenever $c(v_2) \subset c(v_1)$. Also, from (iii) in the construction, there are no two distinct directed paths connecting v_1 and v_2, one of which is an arc. The in–degree of the vertices of $\mathcal{H}$ do not change under the construction and each new vertex has in-degree one. It follows that (b) and (c) hold. □

We end this section by noting that regular hybrids are suitable for a 'temporal representation' in a sense that we will define next.

Let $\mathcal{H}$ be a hybrid phylogeny. We say that $\mathcal{H}$ has a **temporal representation** if there exists a map $f : V(\mathcal{H}) \to \{0, 1, 2, \dots\}$ with the following properties:

(i) For any vertex v of $\mathcal{H}$ with $d^-(v) = 1$ and the arc (u, v) of $\mathcal{H}$, $f(u) < f(v)$.

(ii) For any vertex v of $\mathcal{H}$ with $d^-(v) \geq 2$, and for all arcs (u_i, v) of $\mathcal{H}$ ($i \in \{1, 2, \dots, d^-(v)\}$), $f(u_i) = f(v)$.

Note that a map which satisfies condition $f(u) < f(v)$ for all arcs (u, v) of $\mathcal{H}$ induces an *acycling ordering* on $\mathcal{H}$ and can be defined on any acyclic digraph (see [4]).

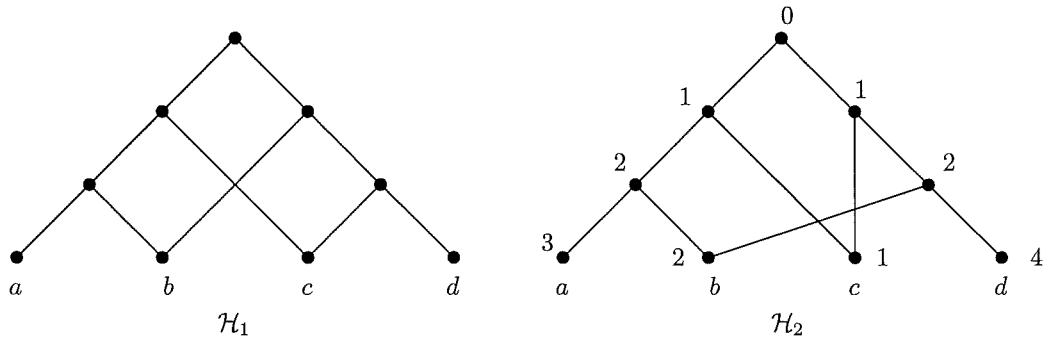


Figure 3.14: Two regular hybrids: $\mathcal{H}_2$ has a temporal representation; $\mathcal{H}_1$ does not.

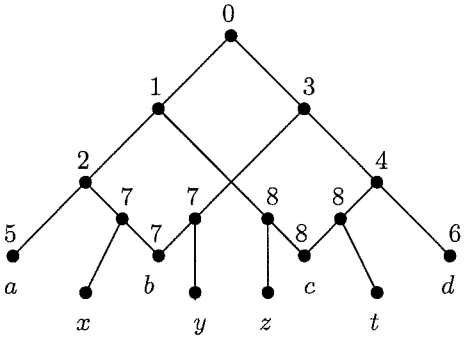


Figure 3.15: A regular hybrid that has a temporal representation, displays $\mathcal{H}_1$ and has the same number of hybrid events.

Proposition 3.5.2. *For any hybrid phylogeny $\mathcal{H}$ there exists a regular hybrid phylogeny $\mathcal{H}'$ with a temporal representation that displays $\mathcal{H}$, and such that $h(\mathcal{H}) = h(\mathcal{H}')$.*

Proof. First, let us observe that without loss of generality, we may assume that $\mathcal{H}$ is a regular hybrid. Then define an acyclic ordering f on $\mathcal{H}$. Construct a new hybrid $\mathcal{H}'$ in the following way: for each vertex v of in–degree greater or equal to two, and for each arc $a_i = (u_i, v)$ of $\mathcal{H}$, $i \in \{1, \ldots, d^-(v)\}$, subdivide the arc a_i by a new vertex w_i, and adjoin a new leaf l_i via a new arc (w_i, l_i). Then $\mathcal{H}'$ is regular, displays $\mathcal{H}$ and $h(\mathcal{H}) = h(\mathcal{H}')$. Define $\overline{f} : V(\mathcal{H}') \to \{0, 1, \ldots\}$ that extends f and such that $\overline{f}(w_i) = \overline{f}(v)$ and $\overline{f}(l_i) > \overline{f}(v)$. The map $\overline{f}$ is a temporal representation for $\mathcal{H}'$. □

3.6 The cluster union hybrid

In this section we will investigate the displaying of a tree by a regular hybrid. We will show that, for any collection $\mathcal{P}$ of rooted phylogenetic trees there is a canonical regular hybrid that displays each of the trees in $\mathcal{P}$, the cluster union hybrid. The particular case when $\mathcal{P}$ consists of two trees is considered.

Let $\mathcal{H} = ((V, A), \psi)$ be a regular hybrid phylogeny on X. If v is a vertex of $\mathcal{H}$ and $V' = \{w \in V(\mathcal{H}) : c(w) \subseteq c(v)\}$, then the subgraph generated by V' together with $\psi|c(v)$ defines a regular hybrid on $c(v)$. We will denote this hybrid by $\mathcal{H}|c(v)$. Note that, for each $v \in V(\mathcal{H})$, $\mathcal{H}$ displays $\mathcal{H}|c(v)$.

Lemma 3.6.1. *Let $\mathcal{H}$ be a regular hybrid phylogeny on X and $\mathcal{T}$ be a rooted phylogenetic X-tree. If $c(\mathcal{T}) \subseteq c(\mathcal{H})$ then $\mathcal{H}$ displays $\mathcal{T}$.*

Proof. The proof is by induction on the height of the tree. Clearly, the statement is true if the height of the tree is 1. Now assume that the proposition holds for any regular hybrid phylogeny and any rooted tree where the height of the latter is at most n. Let $\mathcal{T}$ be a tree of height $n+1$ and denote by ρ and ρ' the roots of $\mathcal{T}$ and $\mathcal{H}$, respectively. We have $c(\rho) = c(\rho') = X$. Let $v_1, v_2, \ldots v_p$ be the vertices of $\mathcal{T}$ that are immediate descendants of ρ and, for each $i \in \{1, 2, \ldots, p\}$, denote by A_i the cluster of $\mathcal{T}$ corresponding to v_i. Then $\{A_i : 1 \leq i \leq p\}$ is a partition of X.

Since $c(\mathcal{T}) \subseteq c(\mathcal{H})$ it follows that, for each i, there exists a vertex u_i of $\mathcal{H}$ with $c(u_i) = A_i$. Let $\mathcal{H}_i$ be the regular hybrid whose set of clusters consists of the subsets of A_i that are clusters of $\mathcal{H}$ and let $\mathcal{T}_i$ be the rooted phylogenetic tree obtained by restricting $\mathcal{T}$ to A_i. Since $\{A_1, A_2, \ldots A_p\}$ partitions X, it follows that for every vertex w that is a descendant of one of the vertices $u_1, u_2, \ldots u_p$, there exists a unique j, $1 \leq j \leq p$, with $c(w) \subseteq A_j$. Thus $\mathcal{H}_i$ is the restriction of $\mathcal{H}$ to A_i, so it is regular. Furthermore, since $c(\mathcal{T}) \subseteq c(\mathcal{H})$, it follows that $c(\mathcal{T}_i) \subseteq c(\mathcal{H}_i)$ for all i. As the height of $\mathcal{T}_i$ is n, by the induction assumption it follows that $\mathcal{H}_i$ displays $\mathcal{T}_i$. To prove that $\mathcal{H}$ displays $\mathcal{T}$, let us observe that for all pairs $i \neq j$, there is no directed path from u_i to u_j in $\mathcal{H}$. Then, for each i, there exists at least one path in $\mathcal{H}$ from the root ρ' to u_i that avoids the vertices $u_1, u_2, \ldots, u_{i-1}, u_{i+1} \ldots u_p$. Hence $\mathcal{H}$ displays $\mathcal{T}$. □

Corollary 3.6.2. *Let $\mathcal{H}$ be a regular hybrid phylogeny on X and $\mathcal{T}$ be a rooted phylogenetic X'-tree with $X' \subseteq X$. If $c(\mathcal{T}) \subseteq c(\mathcal{H})$ then $\mathcal{H}$ displays $\mathcal{T}$.*

Proof. If $c(\mathcal{T}) \subseteq c(\mathcal{H})$ then $X' \in c(\mathcal{H})$ and $\mathcal{T}$ is displayed by $\mathcal{H}|X'$. □

Let $\mathcal{P}$ be a collection of rooted phylogenetic X-trees. We denote by $\mathcal{H}[\mathcal{P}]$ the regular hybrid phylogeny whose set of clusters is $\bigcup_{\mathcal{T}\in\mathcal{P}} c(\mathcal{T})$. As a consequence of Lemma 3.6.1, we obtain the following:

Corollary 3.6.3. *Let $\mathcal{P}$ be a collection of X-trees. Then $\mathcal{H}[\mathcal{P}]$ displays $\mathcal{P}$.*

Consequently, a natural way to obtain a regular hybrid that displays a collection $\mathcal{P}$ of rooted phylogenetic trees with the same set of leaves is by taking the cover hybrid of the union of the sets of clusters of the trees in $\mathcal{P}$. We consider the particular case when $\mathcal{P}$ consists of two rooted phylogenetic trees. We call it the **cluster union hybrid** of the two trees.

Lemma 3.6.4. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted phylogenetic X–trees. Then the following statements hold.*

(i) Each vertex v of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ has the in–degree at most two.

(ii) The hybridization number of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ is equal to

$$|\{v \in V(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]) : d^-(v) = 2\}|.$$

Proof. (i) Assume that there is a vertex v of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ of in–degree at least three. Then there exist at least two immediate ancestors u_1 and u_2 of v, such that $c(u_1)$ and $c(u_2)$ are distinct and are either both clusters of $\mathcal{T}_1$ or both clusters of $\mathcal{T}_2$. Furthermore, $c(v) \subset c(u_1) \cap c(u_2)$, so $c(u_1) \cap c(u_2) \neq \emptyset$. Then either $c(v) \subset c(u_1) \subset c(u_2)$ or $c(v) \subset c(u_2) \subset c(u_1)$. In the former case, it follows from the definition of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ that (u_2, v) is not an arc of the hybrid, a contradiction. Similarly, in the latter case (u_1, v) is not an arc of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$.

(ii) From part (i) it follows that the hybridization vertices of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$) are precisely the vertices of in–degree two. The conclusion is now a consequence of the definition of the hybridization number. □

For two rooted phylogenetic X–trees $\mathcal{T}_1$ and $\mathcal{T}_2$, let φ and f be the mappings of $c(\mathcal{T}_1) \times c(\mathcal{T}_2)$ into $2^{c(\mathcal{T}_1) \cup c(\mathcal{T}_2)}$ defined by

$$\varphi(A, B) = \{S \subseteq A \cap B : S \in c(\mathcal{T}_1) \cup c(\mathcal{T}_2)\},$$

and

$$f(A, B) = \begin{cases} \emptyset & \text{, if } A \cap B \in \{\emptyset, A, B\} \\ \{C \in \varphi(A, B) : C \text{ is maximal }\} & \text{, otherwise.} \end{cases}$$

Proposition 3.6.5. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted phylogenetic X–trees. Then*

$$\bigcup_{(A,B) \in c(\mathcal{T}_1) \times c(\mathcal{T}_2)} f(A, B)$$

is the collection of clusters corresponding to the hybridization vertices of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$.

Proof. Throughout the proof, we denote $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ by $\mathcal{H}$. By Lemma 3.6.4 (ii), it suffices to prove that

$$\{c(v) : v \in V(\mathcal{H}), d^-(v) = 2\} = \bigcup \{f(A, B) : A \cap B \notin \{\emptyset, A, B\}\}.$$

We will establish equality of these two sets by showing set inclusion in both directions. Let $v \in V(\mathcal{H})$ with $d^-(v) = 2$. It follows that v has two immediate ancestors, v_1 and v_2, such that $c(v) \subseteq c(v_1) \cap c(v_2)$, and without loss of generality, $c(v_i) \in \mathcal{C}(\mathcal{T}_i)$ for each $i \in \{1, 2\}$. Also, $c(v) \in \mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$.

Set $S = c(v)$, $A = c(v_1)$, and $B = c(v_2)$. Clearly, $A \cap B \neq \emptyset$ and $A \neq B$. If $A \subset B$ then $c(v) \subset c(v_1) \subset c(v_2)$, contradictory to the definition of $\mathcal{H}$. Similarly, B is not a subset of A, hence $A \cap B \notin \{\emptyset, A, B\}$. Now S is a maximal element of $\{S \subseteq A \cap B : S \in \mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)\}$ under set inclusion. To prove this, suppose that

there exists $S' \in \mathcal{C}(\mathcal{T}_1) \cup \mathcal{C}(\mathcal{T}_2)$, with $S \subset S' \subseteq A \cap B$. Then there is $v' \in V(\mathcal{H})$ such that $S' = c(v')$. It follows that $c(v) \subset c(v') \subset c(v_1)$, a contradiction with the definition of $\mathcal{H}$. Thus if $d^-(v) = 2$, then $c(v) \in \bigcup \{f(A,B) | A \cap B \notin \{\emptyset, A, B\}\}$.

Consider now $S \in f(A,B)$ where $A \in \mathcal{C}(\mathcal{T}_1)$, $B \in \mathcal{C}(\mathcal{T}_2)$, and $A \cap B \notin \{\emptyset, A, B\}$. Then there exist vertices v, v_1, v_2 such that $c(v) = S$, $c(v_1) = A$, and $c(v_2) = B$. As S is a subset of $A \cap B$, there exist paths in $\mathcal{H}$ from v_1 to v and from v_2 to v. By Lemma 3.6.4 (i), it suffices to prove that $d^-(v) \neq 1$. Assume that $d^-(v) = 1$, and denote by v' the immediate ancestor of v in $\mathcal{H}$. Let $S' = c(v')$. Therefore $v_1 \leq v'$ and hence $S' \subseteq A$. Similarly, $S' \subseteq B$ which is contradictory to the choice of S as a maximal cluster included in $A \cap B$. Consequently, $d^-(v) = 2$, hence S belongs to the set $\{c(v) : v \in V(\mathcal{H}), d^-(v) = 2\}$. This completes the proof of the proposition. □

For example, in Fig. 3.16,
$X = \{1,2,3,4,5,6\}$,
$c(T_1) = \{\{1\},\{2\},\{3\},\{4\},\{1,5\},\{2,3\},\{2,3,4\},\{1,2,3,4,5\},X\}$,
$c(T_2) = \{\{1\},\{2\},\{3\},\{4\},\{2,3\},\{4,5\},\{2,3,4\},\{2,3,4,5,6\},X\}$.

We obtain:
$f(\{1,5\},\{4,5\}) = f(\{1,5\},\{2,3,4,5,6\}) = \{\{5\}\}$,
$f(\{2,3,4\},\{4,5\}) = \{\{4\}\}$,
$f(\{1,2,3,4,5\},\{2,3,4,5,6\}) = \{\{2,3,4\},\{4,5\}\}$ and
$f(A,B) = \emptyset$, otherwise.

One can easily remark that $\{1,2,3,4,5\} \cap \{2,3,4,5,6\} = \{2,3,4,5\}$ is not a cluster, but the subsets $\{2,3,4\},\{4,5\}$ are maximal clusters included in $\{2,3,4,5\}$. According to Proposition 3.6.5, $h(\mathcal{H}[\mathcal{T}_1,\mathcal{T}_2]) = |\{\{5\},\{4\},\{2,3,4\},\{4,5\}\}| = 4$.

As the hybridization vertices of $\mathcal{H}[\mathcal{T}_1,\mathcal{T}_2]$ are precisely the vertices of in–degree two, it follows from Proposition 3.6.5 that

$$h(\mathcal{H}[\mathcal{T}_1,\mathcal{T}_2]) = |\bigcup_{(A,B)\in\mathcal{C}(\mathcal{T}_1)\times\mathcal{C}(\mathcal{T}_2)} f(A,B)|.$$

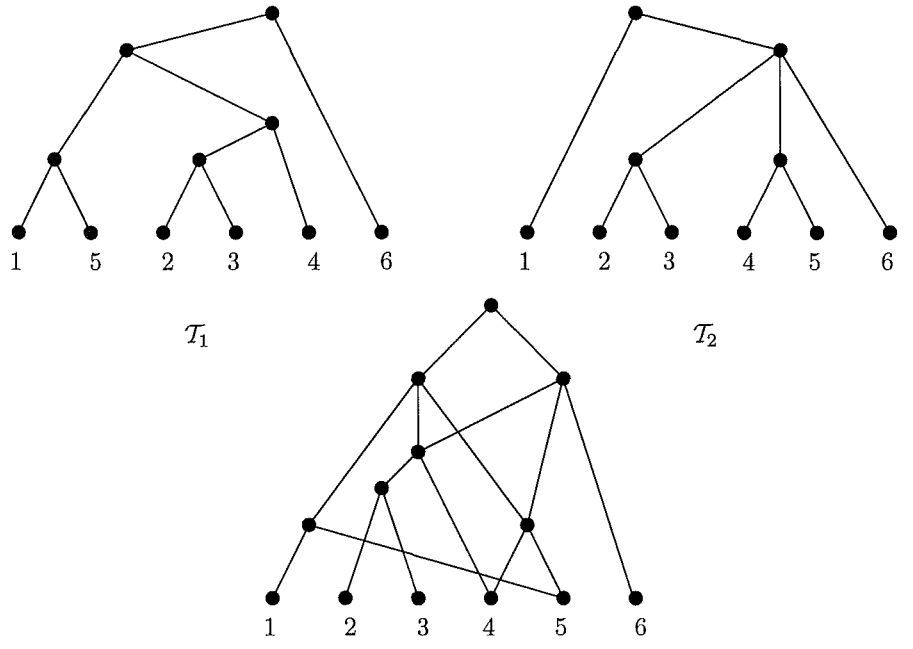


Figure 3.16: The hybrid $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ displays both $\mathcal{T}_1$ and $\mathcal{T}_2$ and $h(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]) = 4$.

Furthermore, a bound on the number of vertices that have in–degree two can be easily obtained as follows. Taking into account that a rooted phylogenetic tree with n leaves has at most $2n - 1$ vertices, such a tree has at most $n - 2$ interior vertices. Therefore the cover hybrid $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ of two rooted phylogenetic trees with the same label set of size n has at most $2(n-2)+n$ vertices different from the root. At least two of these vertices are adjacent to the root, in which case they have the in-degree equal to 1. Hence the number of vertices of $\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$ with in–degree two is at most $2(n-2)+n-2 = 3n-6$. Consequently, $h(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]) \leq 3n-6$.

3.7 The incompatibility graph for a pair of trees

A rooted phylogenetic tree can be reconstructed from its set of clusters [48]. In this section we investigate an extension of this. In particular, for two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$, what information can be inferred about $\mathcal{T}_1$ and $\mathcal{T}_2$ from the union of their cluster sets $c(\mathcal{T}_1) \cup c(\mathcal{T}_2)$? Does the cluster union hybrid associated with the

two trees uniquely determine these trees ? We show that this is the case provided the two trees are sufficiently similar (one can be obtained from the other by a single subtree prune and regraft operation).

Let $\mathcal{C}$ be a collection of subsets of X. The **incompatibility graph** of $\mathcal{C}$ is the graph that has vertex set $\mathcal{C}$ and an edge joining two vertices A and B precisely if $A \cap B \notin \{\emptyset, A, B\}$. The vertices A and B are said to be **incompatible**.

A graph G is said to be *2–colourable* if each vertex of G can be assigned one of two colours so that adjacent vertices are assigned different colours.

We will use the following lemma (see [48]).

Lemma 3.7.1. *Let $\mathcal{C}$ be a cluster system on X. Then there exists a rooted phylogenetic tree $\mathcal{T}$ on X whose set of clusters is $\mathcal{C}$ if and only if, for all $A, B \in \mathcal{C}$,*

$$A \cap B \in \{\emptyset, A, B\}.$$

Moreover, if $\mathcal{C}$ is such a collection, then $\mathcal{T}$ is the unique rooted phylogenetic X–tree having $\mathcal{C}$ as its set of clusters.

Proposition 3.7.2. *Let G be the incompatibility graph of a collection $\mathcal{C}$ of subsets of X. Then G is 2–colourable if and only if there exists a pair of rooted phylogenetic X–trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that $c(\mathcal{T}_1) \cup c(\mathcal{T}_2) = \mathcal{C} \cup X_{triv}$.*

Proof. Assume that G is 2–colourable and let $\mathcal{C}_i$ be the set of vertices of G that are coloured with colour i, $i \in \{1, 2\}$. Then $A \cap B \in \{\emptyset, A, B\}$ whenever both A and B are in $\mathcal{C}_i$. By Lemma 3.7.1, there exists a unique tree $\mathcal{T}_i$ with the cluster set $c(\mathcal{T}_i) = \mathcal{C}_i \cup X_{triv}$.

Conversely, assume that there exists a pair of rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that $c(\mathcal{T}_1) \cup c(\mathcal{T}_2) = \mathcal{C} \cup X_{triv}$. Consider the incompatibility graph G of $\mathcal{C}$ and colour the vertices of G in $c(\mathcal{T}_1)$ one colour and the vertices of G in $c(\mathcal{T}_2)$ another colour. For a vertex in both sets the choice of colour is arbitrary since such a vertex is isolated in G. Let $\{A, B\}$ be an edge of G. Then $A \cap B \notin \{\emptyset, A, B\}$; hence, by Lemma 3.7.1 A and B cannot be clusters of the same tree. So A and B are assigned different colours and this assignment of colours is a 2–colouring of G. □

Proposition 3.7.3. *Let $\mathcal{C}$ be a non-empty collection of clusters on X. Assume that $G = G(\mathcal{C})$ is 2-colourable. Let p be the number of components of G with at least two vertices and let m be the number of isolated vertices of G. Then the number of pairs $(\mathcal{T}_1, \mathcal{T}_2)$ of rooted phylogenetic trees such that $c(\mathcal{T}_1) \cup c(\mathcal{T}_2) = \mathcal{C}$ is equal to $2^{p-1}3^m$ for $p \geq 1$ and $\frac{1}{2}(1+3^m)$ if $p = 0$.*

Proof. Either $p \geq 1$ or $p = 0$. In the former case, the non-isolated vertices of G can be coloured in 2^{p-1} ways. Given a set S with m elements, the number of ordered pairs (A, B) with $A \cup B = S$ equals 3^m, hence the isolated vertices can be coloured in 3^m ways. Therefore, there exist $2^{p-1}3^m$ pairs of trees for any $m \geq 0$.

Now assume that $p = 0$. As in the previous case, we may regard a colouring as an ordered pair (A, B) such that $A \cup B = S$ and $|S| = m$. Since all of the vertices are isolated, we have to identify the pairs (A, B) and (B, A). Taking into account that the only pair (A, B) with $A = B$ is (S, S) there are exactly $\frac{3^m-1}{2} + 1$ pairs. Therefore, there exist $\frac{1+3^m}{2}$ pairs of trees that satisfy the hypothesis. □

Corollary 3.7.4 is an immediate consequence of Proposition 3.7.3.

Corollary 3.7.4. *Let $\mathcal{C}$ be a collection of subsets of X that does not contain any element of X_{triv}, and let G be the incompatibility graph of $\mathcal{C}$. Then there exists exactly one pair $(\mathcal{T}_1, \mathcal{T}_2)$ of rooted phylogenetic trees with $c(\mathcal{T}_1) \cup c(\mathcal{T}_2) = \mathcal{C} \cup X_{triv}$ if and only if either G is a 2-colourable connected graph with at least two vertices or G is the empty graph.*

Example 1. If $\mathcal{C}$ is the empty set then $\mathcal{T}_1 = \mathcal{T}_2$, and their vertices are the root and the leaves.

Example 2. If $|\mathcal{C}| = 1$ then two pairs of trees are obtained (see Figure 3.17).

Example 3. The case $p \geq 1$ is illustrated in Figure 3.18. For $p = 2$ and $m = 0$ there exist two pairs of trees. If we add one isolated vertex, then six pairs of trees are obtained.

Observation. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted phylogenetic trees with the same set of leaves X and let $\mathcal{C}$ be the union of clusters associated with these trees. Denote by*

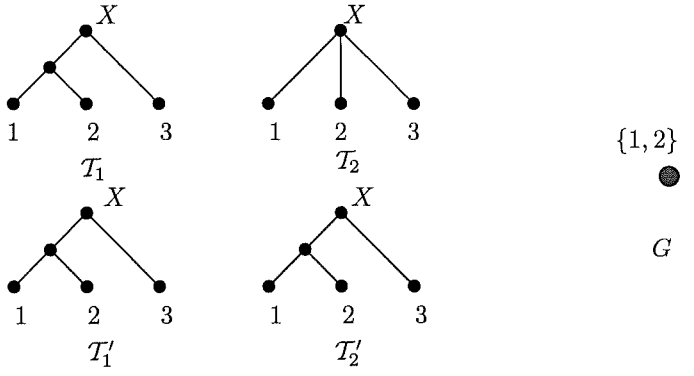


Figure 3.17: Two pairs of trees with the cluster set $\{\{1\},\{2\},\{3\},\{1,2\},\{1,2,3\}\}$ and the incompatibility graph G of $\mathcal{C} = \{1,2\}$.

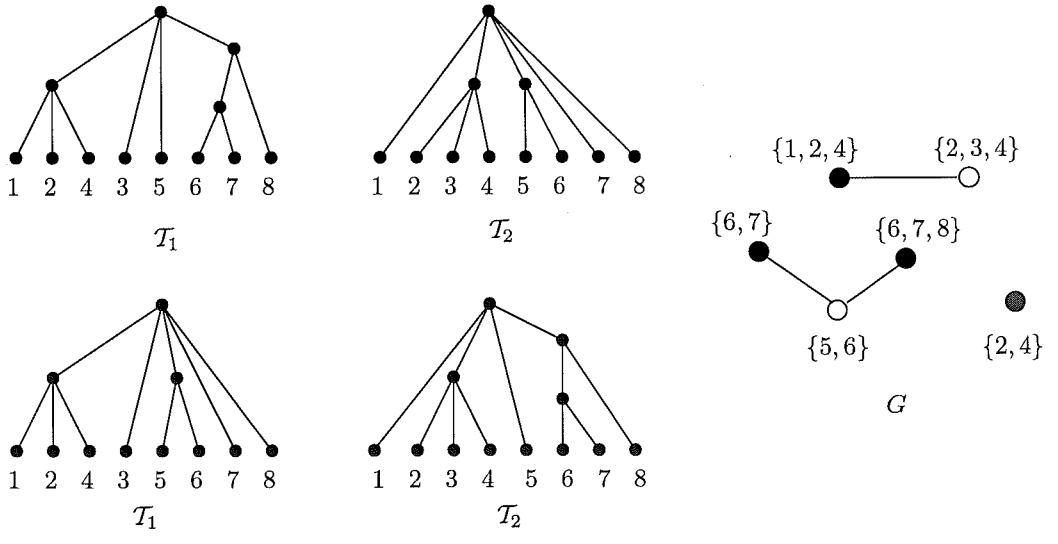


Figure 3.18: The graph G has two components with at least two vertices and one isolated vertex $\{2,4\}$ that can be added to $\mathcal{T}_1$, to $\mathcal{T}_2$, or to both of them.

$\mathcal{H}$ the cluster union hybrid associated to $\mathcal{T}_1$ and $\mathcal{T}_2$. We can use the incompatibility graph to determine the hybridization number $h(\mathcal{H})$.

For every edge of a component with at least two vertices we calculate the intersection of the clusters at the ends of this edge. If this intersection is a cluster in $\mathcal{C}$, then there is a hybridization in the corresponding vertex of $\mathcal{H}$. If all these intersections are clusters, then $h(\mathcal{H})$ is equal to the cardinality of the set of intersections. For the example in Figure 3.18, the set of intersections is $\{\{2,4\},\{6\}\}$, so $h(\mathcal{H}) = 2$.

Lemma 3.7.5. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted phylogenetic trees and denote by $\mathcal{C}_1$ and $\mathcal{C}_2$ the sets of non-trivial clusters of the trees $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. If $A \in \mathcal{C}_1 - \mathcal{C}_2$ then either A is not an isolated vertex of the incompatibility graph G or else $\mathcal{T}_2$ is not binary.*

Proof. If $A \in \mathcal{C}_1 - \mathcal{C}_2$ then A is not a leaf and $A \neq X$. Let $A = \{x_1, \ldots, x_j\}$ and let A' be the minimal cluster of $\mathcal{T}_2$ that contains A. Therefore $A \subseteq A'$ and, since $A \in \mathcal{C}_1 - \mathcal{C}_2$, it follows that $A \neq A'$. As a consequence, there exists at least one leaf $x \in A' - \{x_1, \ldots, x_j\}$. Let B be the direct descendant of A' in $\mathcal{T}_2$ containing x. Since A' is the least common ancestor of $x_1, \ldots, x_j$, A is not included in B. If $A \cap B \neq \emptyset$ then A is not isolated in G.

Therefore we may assume that $A \cap B = \emptyset$. If $\mathcal{T}_2$ is a binary tree, then there exists a direct descendant of A' in $\mathcal{T}_2$ containing $\{x_1, \ldots, x_j\}$ which is contradictory to the definition of A'. □

The next corollary is a straightforward consequence of Proposition 3.7.3 and Lemma 3.7.5.

Corollary 3.7.6. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary trees and let $\mathcal{C} = c(\mathcal{T}_1) \cup c(\mathcal{T}_2)$. If the incompatibility graph $\mathcal{G}$ of $\mathcal{C}$ has at most one component with at least two vertices, then $\mathcal{T}_1$ and $\mathcal{T}_2$ are the only pair of rooted binary phylogenetic trees whose union of clusters is $\mathcal{C}$. Furthermore, if $\mathcal{G}$ consists of isolated vertices, then $\mathcal{T}_1$ and $\mathcal{T}_2$ are isomorphic.*

It follows from Corollary 3.7.6 that two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ for which the incompatibility graph $\mathcal{G}$ has at most one component with at least

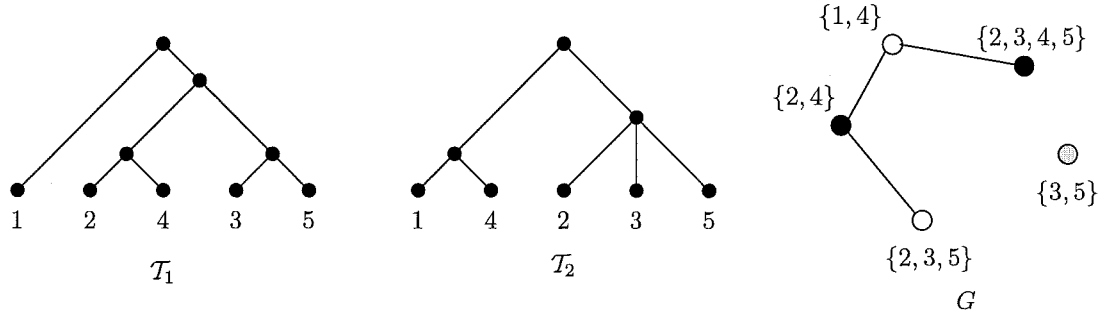


Figure 3.19: The cluster 35 is an isolated vertex for the incompatibility graph G.

two vertices can be reconstructed from $\mathcal{G}$. More precisely, the clusters of $\mathcal{T}_1$ and $\mathcal{T}_2$ can be constructed as follows. If $\mathcal{G}$ consists of isolated vertices, then $\mathcal{T}_1$ and $\mathcal{T}_2$ are isomorphic, and the vertex set of $\mathcal{G}$ is the set of clusters of $\mathcal{T}_1$. If $\mathcal{G}$ has exactly one component C with at least two vertices, 2–colour this component, and for each $i \in \{1,2\}$, let $\mathcal{C}_i$ be the union of the set of isolated vertices of $\mathcal{G}$ and the subset of vertices of $\mathcal{C}$ assigned one particular colour. Then, for each i, $\mathcal{C}_i$ is the set of clusters of $\mathcal{T}_i$.

Proposition 3.7.7. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two distinct rooted binary phylogenetic trees on X, such that $\mathcal{T}_2$ can be obtained from $\mathcal{T}_1$ by a single rSPR operation. Then:*

(i) The incompatibility graph of $\mathcal{C} = c(\mathcal{T}_1) \cup c(\mathcal{T}_2)$ has exactly one component with at least two vertices.

(ii) Both $\mathcal{T}_1$ and $\mathcal{T}_2$ can be reconstructed from the incompatibility graph of $c(\mathcal{T}_1) \cup c(\mathcal{T}_2)$.

Proof. Consider a rooted binary tree $\mathcal{T}$ and the tree $\mathcal{T}'$ obtained from it by a single rSPR operation that prunes the subtree with the leaf set A, and reattach it up to the root of the subtree with the leaf set E. The two trees are drawn in Figure 3.20. Note that $c(\mathcal{T}_1) \cap c(\mathcal{T}_2)$ consists of all the clusters of $\mathcal{T}_1$ whose associated vertices do not lie on the path from A to E, except for the vertex V that is the most recent common ancestor of A and E. Let $\mathcal{D}$ be the incompatibility graph of the set of

clusters $c(\mathcal{T}_1)\triangle c(\mathcal{T}_2)$. To prove (i) we will show that the incompatibility graph $\mathcal{G}$ of $\mathcal{D}$ consists of a single component. The relationship between clusters is easily checked and the pairs of incompatible vertices are:

D_i, D $(0 \leq i \leq k)$; $D_i, B_l \cup A$ $(0 \leq i \leq k,\ 1 \leq l \leq j)$;
$U, D_i - A$, $(0 \leq i \leq k)$; $D_l - A, D_i$ $(1 \leq i \leq k-1,\ i+1 \leq l \leq k)$;
D, B_i $(1 \leq i \leq j)$; $B_l \cup A, B_i$ $(1 \leq l \leq j-1,\ l+1 \leq i \leq j)$.

The incompatibility graph of $\mathcal{D}$ is drawn in Figure 3.21.

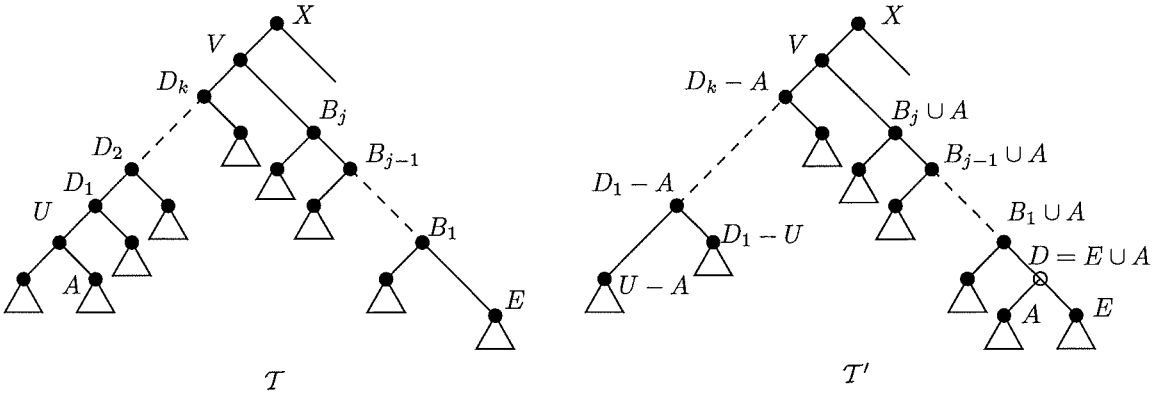


Figure 3.20: The clusters of $\mathcal{T}$ have been changed. The clusters $D_1, \ldots, D_k$ become $D_1 - A, \ldots, D_k - A$ and $B_1, \ldots B_j$ are replaced by $B_1 \cup A, \ldots, B_j \cup A$.

In a similar way, one can prove the result for the other two types of rSPR operations. This completes the proof of (i).

To prove (ii) we notice that, as (i) holds, we can apply the construction described after Corollary 3.7.6 to obtain the two trees. □

Corollary 3.7.8. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary phylogenetic trees on X, such that $d_{rSPR}(\mathcal{T}_1, \mathcal{T}_2) = 1$. Then*

$$h(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]) = \frac{1}{2}|c(\mathcal{T}_1)\triangle c(\mathcal{T}_2)| = k + j + 1.$$

Proof. Suppose that $\mathcal{T}_2$ is obtained from $\mathcal{T}_1$ by a single rSPR operation as shown in Figure 3.20. Denote by $\mathcal{C} = c(\mathcal{T}_1) \cup c(\mathcal{T}_2)$. Let us observe that if A and B are

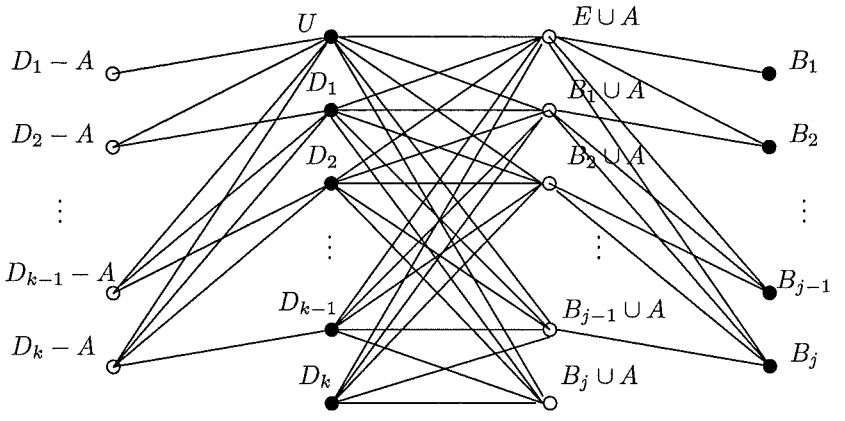


Figure 3.21: The incompatibility graph for the trees $\mathcal{T}_1$ and $\mathcal{T}_2$ has only one component with at least two vertices. The vertices which are not contained in this component (are isolated vertices for the incompatibility graph) belong to both $\mathcal{T}_1$ and $\mathcal{T}_2$.

clusters in $\mathcal{C}$ associated to incompatible pairs of vertices, then $A \cap B \in \mathcal{C}$. It follows that

$$h(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]) = |\{A, E, U - A, D_1 - A, \ldots, D_{k-1} - A, B_1, B_2, \ldots, B_{j-1}\}| = k + j + 1.$$

For the other types of rSPR operations, the hybridization number can be calculated in a similar way. □

We end this section by noting that Proposition 3.7.7 cannot be extended to two rooted binary phylogenetic trees that need more than one rSPR operation to be obtained one from the other. To see this, consider the two rooted binary phylogenetic trees shown in Figure 3.22. Two rSPR operations are required to obtain $\mathcal{T}_2$ from $\mathcal{T}_1$, respectively $\mathcal{T}_2'$ from $\mathcal{T}_1'$. On the other hand,

$$c(\mathcal{T}_1) \cup c(\mathcal{T}_2) = c(\mathcal{T}_1') \cup c(\mathcal{T}_2').$$

In particular, the associated incompatibility graphs are identical.

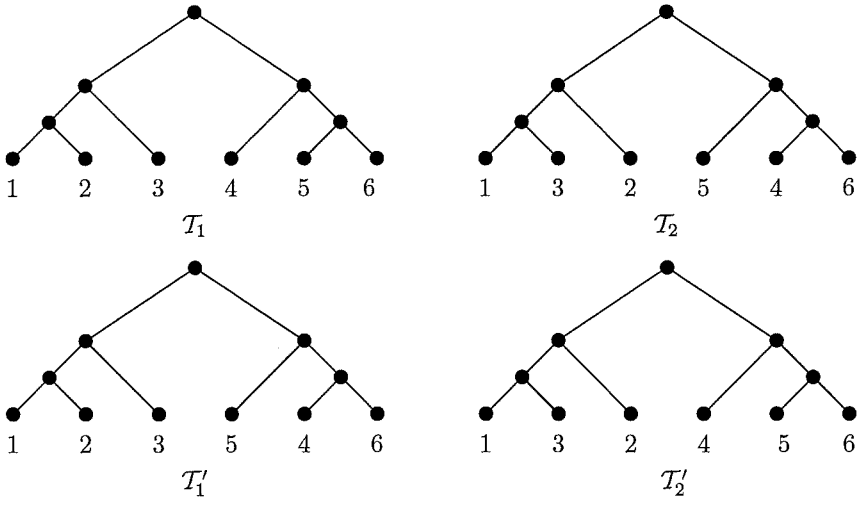


Figure 3.22: Two pairs of trees with the same union of clusters.

3.8 Some remarks on the notion of 'display'

We end this chapter with some remarks regarding possible variants of the notion of 'display', particularly for the case of a hybrid that displays a tree. According to the definition used in this thesis, the hybrid $\mathcal{H}$ drawn in Figure 3.23 displays neither $\mathcal{T}_1$ nor $\mathcal{T}_2$. The problem appears because the paths from the root of $\mathcal{T}_1$ (or $\mathcal{T}_2$) to the leaves labelled by b, respectively by c, cannot be 'displayed' by $\mathcal{H}$ as they need to 'share' the vertex v.

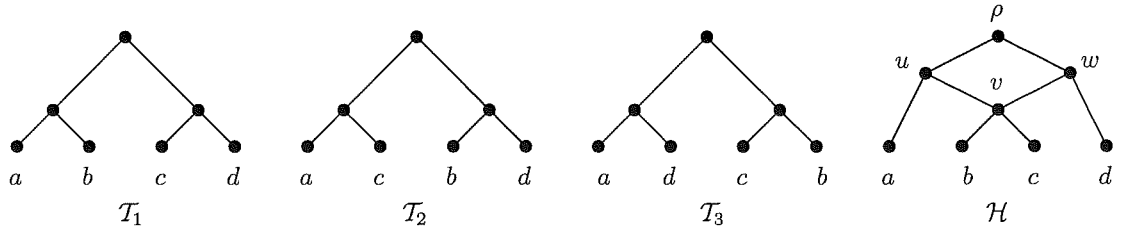


Figure 3.23: The hybrid $\mathcal{H}$ 'vertex–sharing' displays both $\mathcal{T}_1$ and $\mathcal{T}_2$.

Let $\mathcal{H}$ be a hybrid phylogeny on X, ω a vertex of $\mathcal{H}$ with $X' \subseteq c(\omega)$. Denote by $P(\omega)$ the set of paths starting in ω and ending in a leaf of $\mathcal{H}$. Consider $\mathcal{P} \subseteq \mathcal{P}(\omega)$

such that for any $x \in X'$, there exists a unique path in $\mathcal{P}$ that ends in x. Let

$$V(\mathcal{P}) = \{v \in V(\mathcal{H}) : v \text{ is a vertex of } p \text{ for some } p \in \mathcal{P}\}.$$

If $p \in \mathcal{P}$, $p = \omega v_1 v_2 \dots v_n$ and $s = v_i$ for some i, $1 \leq i \leq n$, denote by $pr(s,p)$ the path $\omega v_1 v_2 \dots v_{i-1}$ and consider the set

$$\mathcal{P}_r(s) = \{pr(s,p) : p \in \mathcal{P}\}.$$

Construct a new graph having:

- the set of vertices: ω, $(s, pr(s,p))$, $p \in \mathcal{P}$ (For any vertex s of $V(\mathcal{P})$, correspond $|\mathcal{P}_r(s)|$ vertices.)
- the set of arcs: there is an arc from $(s_1, pr(s_1,p))$ to $(s_2, pr(s_2,p))$ precisely if $pr(s_2,p) = pr(s_1,p)s_1$.

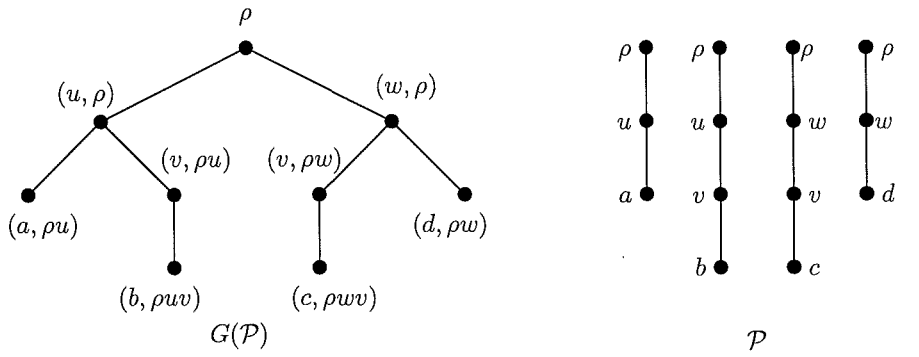


Figure 3.24: A collection of paths $\mathcal{P}$ and the graph $G(\mathcal{P})$ to show that the hybrid $\mathcal{H}$ in Figure 3.23 'vertex–sharing' displays $\mathcal{T}_1$.

We denote this graph by $G(\mathcal{P})$ and we call it the **equivalent–path graph**. For an example, see Figure 3.24.

Note that the above construction is equivalent with the following. Consider the equivalence relation on $V(\mathcal{P}) \times \mathcal{P}$ given by

$$(u,p) \sim (u',p') \text{ if and only if } \begin{cases} u = u' = \omega \\ \text{or} \\ u = u' \text{ and } (da(u),p) \sim (da(u'),p') \end{cases}$$

where $da(u)$ denotes the direct ancestor of u on the path p. For $(v,p) \in V(\mathcal{P}) \times \mathcal{P}$, denote by $[v,p]$ the equivalence class of (v,p); that is $[v,p] = \{(v',p') : (v',p') \sim (v,p)\}$. Then construct the graph $G(\mathcal{P})$ as follows:

- the set of vertices: $\{[v,p] : (v,p) \in V(\mathcal{P}) \times \mathcal{P}\}$
- the set of arcs: there is an arc between $[(v,p)]$ and $[(v',p')]$ if and only if there is $q \in \mathcal{P}$ such that $(v,p) \sim (v,q)$, $(v',p') \sim (v',q)$ and (v,v') is an arc of q.

Now we can introduce a variant of display that allows 'vertex-sharing'.

Let $\mathcal{H}$ be a hybrid phylogeny on X and $\mathcal{T}$ a rooted phylogenetic X'-tree with $X' \subseteq X$. We say that $\mathcal{H}$ **displays** $\mathcal{T}$ if there exists $\omega \in V(\mathcal{H})$ with $X' \subseteq c(\omega)$ and there exists a collection of paths $\mathcal{P}$ corresponding to ω and X' such that $G(\mathcal{P})$ is a refinement of $\mathcal{T}$.

Note that, according to the above definition, the hybrid $\mathcal{H}$ drawn in Figure 3.23 displays $\mathcal{T}_1$ and $\mathcal{T}_2$ but does not display $\mathcal{T}_3$. Also, let us observe that, if $|V(\mathcal{P})| = |V(G(\mathcal{P}))|$, we obtain the definition introduced in Section 3.4 and used in this thesis.

3.9 Some questions for future work

Regular hybrids have an interesting combinatorial structure that gives rise to many further questions, for example, find the number of regular hybrids with n leaves and exactly (respectively at most) k hybridization events.

Besides its mathematical interest, investigating the structure of regular hybrids might have some useful application to biological problems. For example, we have proved that each hybrid phylogeny can be modified to obtain a regular hybrid by adding leaves. This operation has a biological meaning—the new leaves correspond to species that may have been involved in the evolution in the past or have yet to be sampled. So an important question is this: For a given hybrid, what is the minimum number of leaves that should be added to transform it into a regular one?

Also, given a (regular) hybrid, find the minimum number of leaves that need to be added in order to obtain a temporal representation as described in Section 3.5.

It would also be interesting to investigate the mathematical properties of the more general notion of display defined in Section 3.8.

Chapter 4

Measuring the dissimilarities between trees

In this chapter, we use the formalism introduced in Chapter 3 to study the following problem: *given a collection of rooted phylogenetic trees, how can these trees be displayed by a single hybrid phylogeny with a minimum number of hybrid events?* Apart from its mathematical interest, there is a biological motivation for considering this problem. This motivation has been summarized in [42]:

> *... it would be more useful to have analytical methods that allowed biologists to visualise the phylogenetic relationships between plant species (not just between genes) and to have methods that could account for patterns in the data arising from species hybridisation events. Ideally such methods would also allow researchers to infer the minimum number of hybridisation events needed to explain extant diversity.*

In general, the cluster union hybrid (although it is a 'canonical' one) has more hybrid events than are required. On the other hand, the number of hybrid events can be reduced (in some cases greatly reduced, as we will show later in this section) if other species (not sampled by any of the input trees) are permitted. This phenomenon is biologically motivated since other species (including ones that are

extinct now) may have been involved in evolution in the past. These aspects are illustrated in the following example. The three hybrids shown in Figure 4.1 display both $\mathcal{T}_1$ and $\mathcal{T}_2$. Note that $\mathcal{H}_1$ is the cluster union hybrid associated with $\mathcal{T}_1$ and $\mathcal{T}_2$ and $h(\mathcal{H}_1) = 6$. The hybrid $\mathcal{H}_3$ has only two hybrid events but one extra–leaf x. If we use the leaf 6 rather then an extra–leaf, then we obtain $\mathcal{H}_2$. The set of leaves is preserved but $h(\mathcal{H}_2) = 3$.

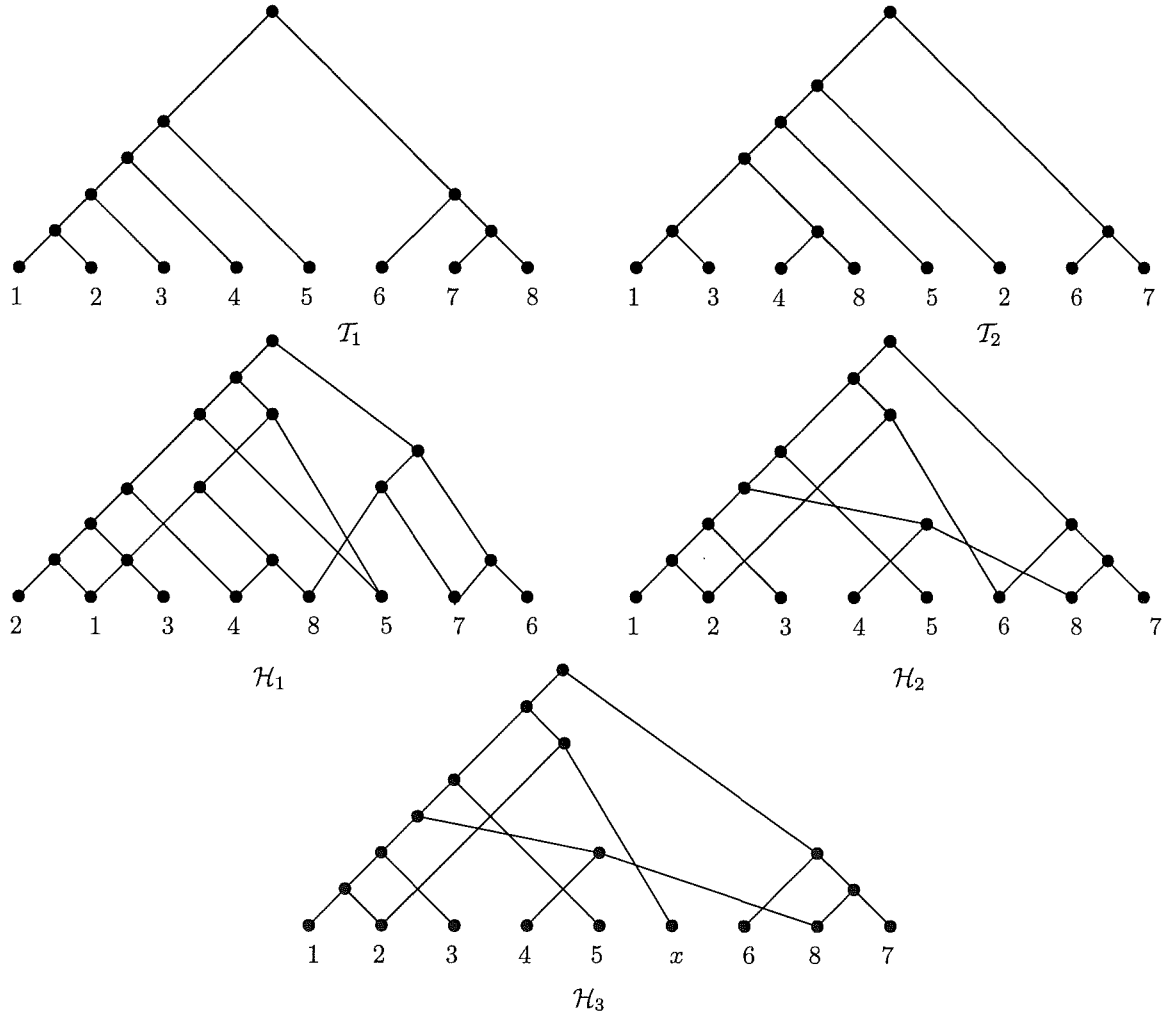


Figure 4.1: Two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ and three regular hybrid phylogenies $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$ that display both trees.

We will show that the minimum number of hybrid events required to display two rooted binary phylogenetic trees by the same hybrid is bounded below by the rSPR distance between the two trees. We also examine situations when equality holds. Results in this chapter will play an important role in Chapter 5.

4.1 Displaying trees with a minimum number of hybrid events

Let $\mathcal{P}$ be a collection of rooted phylogenetic trees and let

$$\mathcal{L}(\mathcal{P}) = \cup_{\mathcal{T}\in\mathcal{P}}\mathcal{L}(\mathcal{T}).$$

We say that a hybrid $\mathcal{H}$ **displays** $\mathcal{P}$ if each tree of $\mathcal{P}$ is displayed by $\mathcal{H}$. We are interested in displaying $\mathcal{P}$ with a minimum number of hybrid events. This number depends on the conditions required for the hybrid $\mathcal{H}$ that displays $\mathcal{P}$. For example, if $\mathcal{H}$ is a regular hybrid with $\mathcal{L}(\mathcal{P}) \subseteq \mathcal{L}(\mathcal{H})$, it follows from Proposition 3.5.1 that the restriction to regular hybrid phylogenies does not change the minimum number. However, this is not the case when we do not allow extra–leaves.

We now introduce three possible measures of hybridization. For a collection $\mathcal{P}$ of rooted phylogenetic trees we define:

$$\begin{aligned} h(\mathcal{P}) &= \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a hybrid that displays } \mathcal{T} \text{ and } \mathcal{T}' \text{ with } \mathcal{L}(\mathcal{H}) = \mathcal{L}(\mathcal{P})\}, \\ h_r(\mathcal{P}) &= \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a regular hybrid that displays } \mathcal{P} \text{ and } \mathcal{L}(\mathcal{H}) = \mathcal{L}(\mathcal{P})\}, \\ h_r^+(\mathcal{P}) &= \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a regular hybrid that displays } \mathcal{P} \text{ and } \mathcal{L}(\mathcal{P}) \subseteq \mathcal{L}(\mathcal{H})\}. \end{aligned}$$

If $\mathcal{P} = \{\mathcal{T}, \mathcal{T}'\}$, we will write $h(\mathcal{T}, \mathcal{T}')$ instead of $h(\{\mathcal{T}, \mathcal{T}'\})$. Similarly for h_r and h_r^+. Let us observe that, as the following proposition shows, no generality is lost in restricting our discussion to rooted binary phylogenetic trees with the same set of leaves.

Proposition 4.1.1. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary phylogenetic trees and denote by $\mathcal{L}_{12} = \mathcal{L}(\mathcal{T}_1) \cap \mathcal{L}(\mathcal{T}_2)$. Let*

$$h(\mathcal{T}_1, \mathcal{T}_2) = \min\{h(\mathcal{H}) : \mathcal{H} \textit{ is a hybrid that displays } \mathcal{T}_1 \textit{ and } \mathcal{T}_2\}.$$

Then

$$h(\mathcal{T}_1, \mathcal{T}_2) = h(\mathcal{T}_1|\mathcal{L}_{12}, \mathcal{T}_2|\mathcal{L}_{12}).$$

Proof. By Lemma 3.4.2 we have that $h(\mathcal{T}_1|\mathcal{L}_{12}, \mathcal{T}_2|\mathcal{L}_{12}) \leq h(\mathcal{T}_1, \mathcal{T}_2)$.

Let $\mathcal{H}$ be a hybrid that displays $\mathcal{T}_1|\mathcal{L}_{12}$ and $\mathcal{T}_2|\mathcal{L}_{12}$. From the definition of displaying, it follows that $\mathcal{L}_{12} \subseteq \mathcal{L}(\mathcal{H})$ and, for each $i \in \{1, 2\}$ there exists a rooted subdigraph (tree) $\mathcal{H}_i$ of $\mathcal{H}$ that is a refinement of $\mathcal{T}_i|\mathcal{L}_{12}$. Now consider the root ρ of $\mathcal{H}$ as a pendant vertex attached to the original root. Then we can consider the root of $\mathcal{H}_i$ as ρ and adjoin the vertices labelled by $\mathcal{L}(\mathcal{T}_i) - \mathcal{L}_{12}$ to edges of $\mathcal{H}_i$ such that the obtained tree is a refinement of $\mathcal{T}_i$. The digraph obtained by this construction is a hybrid that displays both $\mathcal{T}_1$ and $\mathcal{T}_2$ and has the same hybridization number as $\mathcal{H}$. Therefore $h(\mathcal{T}_1, \mathcal{T}_2) \leq h(\mathcal{T}_1|\mathcal{L}_{12}, \mathcal{T}_2|\mathcal{L}_{12})$. □

It is straightforward to observe that for all rooted phylogenetic X–trees

$$\max\{h(\mathcal{T}, \mathcal{T}'), h_r^+(\mathcal{T}, \mathcal{T}')\} \leq h_r(\mathcal{T}, \mathcal{T}').$$

Also, $h(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}', \mathcal{T})$, and similar relations hold for h_r and h_r^+. Clearly, $h(\mathcal{T}, \mathcal{T}) = h_r(\mathcal{T}, \mathcal{T}) = h_r^+(\mathcal{T}, \mathcal{T}) = 0$ but in general, $h(\mathcal{T}, \mathcal{T}') = 0$ does not imply $\mathcal{T} = \mathcal{T}'$. To see this, consider two non-isomorphic rooted X-trees $\mathcal{T}$ and $\mathcal{T}'$ such that $\mathcal{T}'$ is a refinement of $\mathcal{T}$. Then $\mathcal{T}'$ itself is a hybrid that displays both $\mathcal{T}$ and $\mathcal{T}'$, so $h(\mathcal{T}, \mathcal{T}') = 0$. The corresponding implications for h_r and h_r^+ do not hold too. Nevertheless, if we consider only binary rooted phylogenetic X–trees, then $h(\mathcal{T}, \mathcal{T}') = 0$ if and only if $\mathcal{T} = \mathcal{T}'$. A similar result holds for h_r and h_r^+. In other words, h, h_r and h_r^+ are dissimilarity maps on the collection of rooted binary X–trees. However, as we will show later, these maps do not satisfy the triangle inequality.

The result in the following proposition shows that displaying two phylogenetic X–trees by a hybrid with the same set X of leaves is equivalent to displaying trees by a regular hybrid whose set of leaves contains X.

Proposition 4.1.2. *For any two rooted binary phylogenetic X–trees $\mathcal{T}_1$ and $\mathcal{T}_2$,*

$$h_r^+(\mathcal{T}_1, \mathcal{T}_2) = h(\mathcal{T}_1, \mathcal{T}_2).$$

Proof. By Proposition 3.5.1 it follows that $h_r^+(\mathcal{T}_1, \mathcal{T}_2) \leq h(\mathcal{T}_1, \mathcal{T}_2)$.

We now prove the inequality $h(\mathcal{T}_1, \mathcal{T}_2) \leq h_r^+(\mathcal{T}_1, \mathcal{T}_2)$. Let $\mathcal{H}^+$ be a regular hybrid that displays $\mathcal{T}_1$ and $\mathcal{T}_2$, and such that $X \subseteq \mathcal{L}(\mathcal{H}^+)$. Then for each $i \in \{1, 2\}$, there exists a rooted subdigraph $\mathcal{H}_i$ of $\mathcal{H}^+$ that is a refinement of $\mathcal{T}_i$. It follows that $\mathcal{H}_i$ is a tree with the leaf set labelled by X. Consider now the subdigraph $\mathcal{H}$ of $\mathcal{H}^+$ that is obtained by restricting $\mathcal{H}^+$ to the vertices and arcs that are used by either of $\mathcal{H}_i$, $i \in \{1, 2\}$. By possibly adding a new vertex that is joined to each of the vertices of $\mathcal{H}$ of in–degree zero, we may assume that $\mathcal{H}$ is a rooted digraph. Furthermore, all the vertices of out–degree zero of $\mathcal{H}$ are labelled by an element of X, for otherwise, there exists a vertex in $\mathcal{H}$ that is not induced by a vertex of $\mathcal{H}_i$. It follows that, up to suppressing degree-two vertices, $\mathcal{H}$ is a hybrid with $\mathcal{L}(\mathcal{H}) = X$ that displays $\mathcal{T}_1$ and $\mathcal{T}_2$. Since each non-root vertex v of $\mathcal{H}$ corresponds to a vertex of $\mathcal{H}^+$, and the set of arcs directed towards v is a subset of the arcs directed towards the corresponding vertex in $\mathcal{H}^+$, it follows that $h(\mathcal{H}) \leq h(\mathcal{H}^+)$. Thus $h(\mathcal{T}_1, \mathcal{T}_2) \leq h_r^+(\mathcal{T}_1, \mathcal{T}_2)$. □

As a consequence of Proposition 4.1.2 and previous observations, for each pair of rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$, we have that:

$$h(\mathcal{T}_1, \mathcal{T}_2) = h_r^+(\mathcal{T}_1, \mathcal{T}_2) \leq h_r(\mathcal{T}_1, \mathcal{T}_2).$$

We will prove that the above inequality is strict. Moreover, for a pair of rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$, the difference between $h_r(\mathcal{T}_1, \mathcal{T}_2)$ and $h_r^+(\mathcal{T}_1, \mathcal{T}_2)$ can be arbitrarily large as the next proposition shows.

Proposition 4.1.3. *For all $n \geq 3$, there exist two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that $h_r^+(\mathcal{T}_1, \mathcal{T}_2) = 1$ and $h_r(\mathcal{T}_1, \mathcal{T}_2) = n - 2$.*

Proof. Let $X = \{1, 2, \ldots, n\}$ and consider the rooted phylogenetic X–trees $\mathcal{T}_1$ and $\mathcal{T}_2$ shown in Figure 4.2. Since both $\mathcal{T}_1$ and $\mathcal{T}_2$ are binary, and neither $\mathcal{T}_1$ nor $\mathcal{T}_2$ displays the other, any regular hybrid that displays both $\mathcal{T}_1$ and $\mathcal{T}_2$ requires at least one hybrid event.

Let $\mathcal{H}^+$ be the hybrid shown in Figure 4.2, where $x \notin X$. Since the hybrid $\mathcal{H}^+$ is regular, displays both $\mathcal{T}_1$ and $\mathcal{T}_2$ and $h(\mathcal{H}^+) = 1$, it follows that $h_r^+(\mathcal{T}_1, \mathcal{T}_2) = 1$.

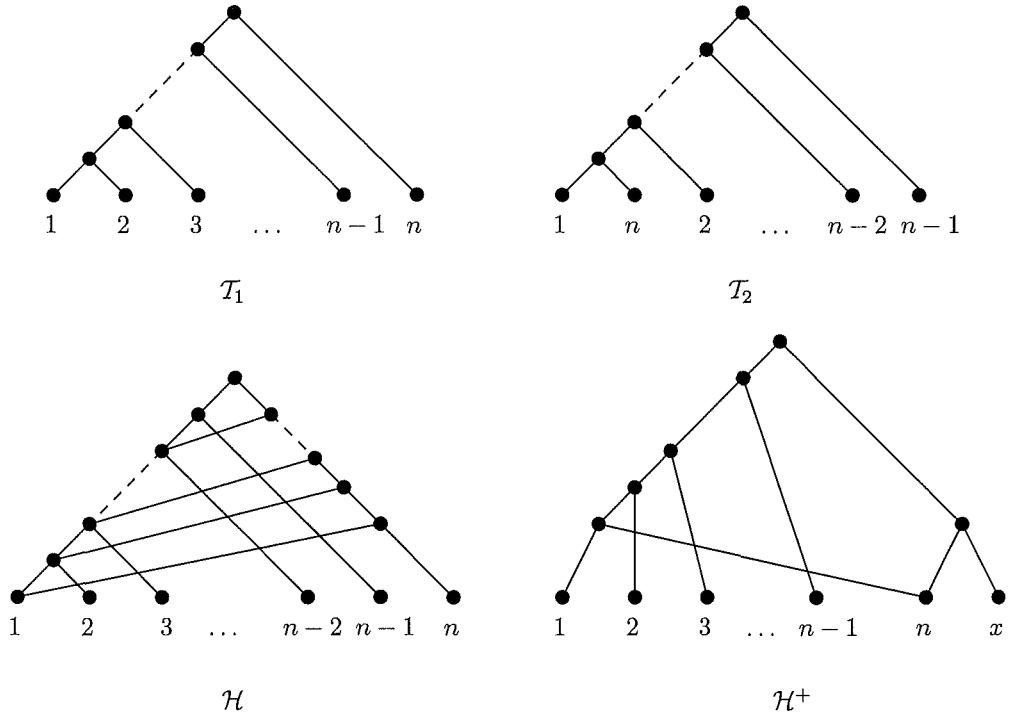


Figure 4.2: Two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ and two regular hybrids that display both $\mathcal{T}_1$ and $\mathcal{T}_2$. Note that $\mathcal{H}$ has the same leaf set. The hybrid $\mathcal{H}^+$ has one extra leaf.

We next show that $h_r(\mathcal{T}_1, \mathcal{T}_2) = n - 2$. Let $\mathcal{H}$ be the regular hybrid as shown in Figure 4.2. Then $\mathcal{H}$ displays $\mathcal{T}_1$ and $\mathcal{T}_2$ and $h(\mathcal{H}) = n - 2$. It remains to prove that $n - 2$ is the minimum hybridization number for all the regular hybrids on X displaying both $\mathcal{T}_1$ and $\mathcal{T}_2$. Let $\mathcal{G}$ be a regular hybrid on $\{1, 2, \ldots, n\}$ that displays $\mathcal{T}_1$. Since G displays $\mathcal{T}_1$, it follows by Lemma 3.4.1 that there exists a strictly increasing map $\varphi : c(\mathcal{T}_1) \to c(\mathcal{G})$ that preserves the leaves. Then

$$\{1\} = \varphi(\{1\}) \subset \varphi(\{1, 2\}) \subset \cdots \subset \varphi(\{1, 2, \ldots, n-1\} \subset \varphi(\{1, 2, \ldots, n-1, n\}).$$

Since $\mathcal{T}_1$ and $\mathcal{G}$ have the same label set, $\varphi(\{1, 2, \ldots, n-1, n\}) = \{1, 2, \ldots, n-1, n\}$. This now implies that $\varphi(\{1, 2, \ldots, i\}) = \{1, 2, \ldots, i\}$, for all i. Therefore $c(\mathcal{T}_1) \subseteq \mathcal{G}$. Similarly, $c(\mathcal{T}_2) \subseteq \mathcal{G}$. It follows that $|V(\mathcal{G})| \geq n + 1 + 2(n - 2) = 3n - 3$.

By Lemma 3.3.5, this implies that

$$h(\mathcal{G}) \geq |V(\mathcal{G})| - 2n + 1 \geq (3n - 3) - 2n + 1 = n - 2.$$

Consequently, $h_r(\mathcal{T}_1, \mathcal{T}_2) = n - 2$. □

Let $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots \mathcal{T}_k\}$ be a set of rooted binary phylogenetic trees with $\mathcal{L}(\mathcal{T}_1) = \mathcal{L}(\mathcal{T}_2) = \cdots = \mathcal{L}(\mathcal{T}_k) = \mathcal{L}$ and for $U \subseteq \mathcal{L}$, let $\mathcal{P}|U = \{\mathcal{T}_1|U, \mathcal{T}_2|U, \ldots \mathcal{T}_k|U\}$. The following proposition is a consequence of the definition of h_r^+ and Proposition 3.4.2.

Proposition 4.1.4. *Let $\mathcal{P}$ be a collection of rooted phylogenetic trees. Then, for all subsets U of $\mathcal{L}(\mathcal{P})$, we have $h_r^+(\mathcal{P}|U) \leq h_r^+(\mathcal{P})$.*

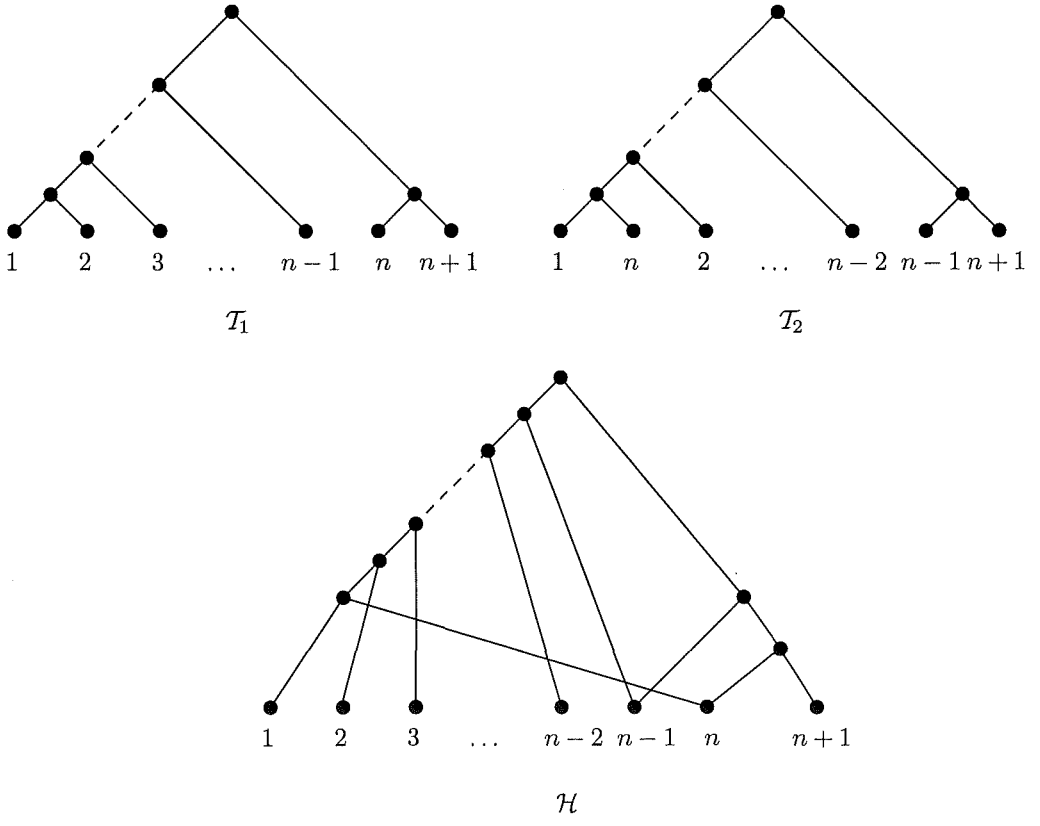


Figure 4.3: Two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ for the proof of Proposition 4.1.5.

Note that the inequality in Proposition 4.1.4 is no longer valid when h_r^+ is replaced by h_r. More precisely, we have the following proposition.

Proposition 4.1.5. *For all integers $n \geq 4$, there exist X, a collection $\mathcal{P}$ of rooted binary phylogenetic trees on X, and a subset U of X such that*

$$h_r(\mathcal{P}|U) \geq h_r(\mathcal{P}) + n - 4.$$

Proof. Let $U = \{1, 2, \ldots, n\}$ be a subset of X, and $P = \{\mathcal{T}_1, \mathcal{T}_2\}$ as shown in Figure 4.3. Then $P|U$ consists of the two rooted binary phylogenetic trees shown in Figure 4.2.

By Proposition 4.1.3, $h_r(P|U) = n - 2$. Furthermore, the hybrid $\mathcal{H}$ shown in Figure 4.3 is regular and displays both $\mathcal{T}_1$ and $\mathcal{T}_2$. Since $h(\mathcal{H}) = 2$, it follows that $h_r(\mathcal{P}) \leq 2$. The proposition now follows. □

We will show now that h_r does not verify the triangle inequality.

Proposition 4.1.6. *There exist three rooted phylogenetic trees $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ such that $h_r(\mathcal{T}_1, \mathcal{T}_3) > h_r(\mathcal{T}_1, \mathcal{T}_2) + h_r(\mathcal{T}_2, \mathcal{T}_3)$.*

Proof. For the trees shown in Figure 4.4, $\mathcal{H}_{i,j}$ is a minimal regular hybrid that displays $\mathcal{T}_i$ and $\mathcal{T}_j$, $i, j \in \{1, 2, 3\}$. It follows that $h_r(\mathcal{T}_1, \mathcal{T}_2) = h_r(\mathcal{T}_2, \mathcal{T}_3) = 1$ and $h_r(\mathcal{T}_1, \mathcal{T}_3) = 3$. □

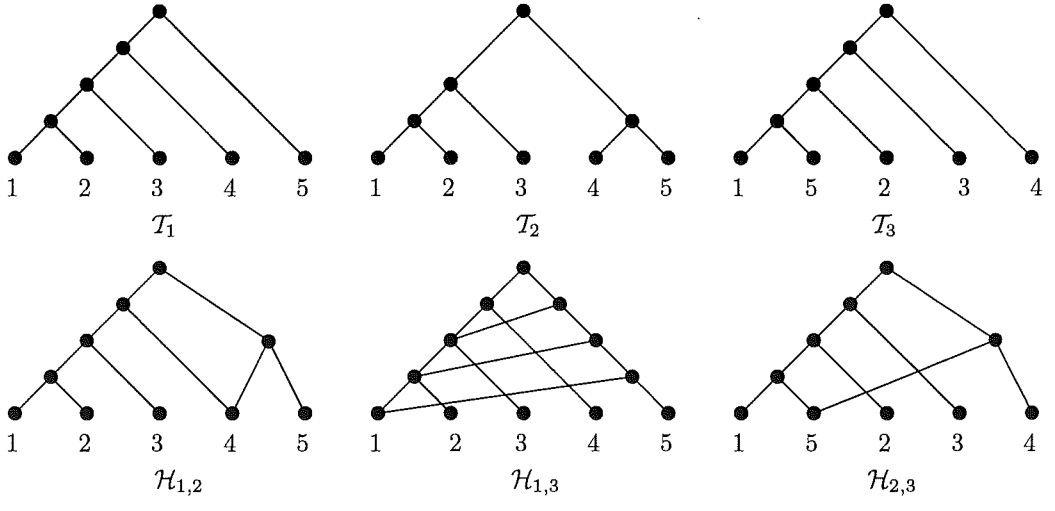


Figure 4.4: Three rooted binary phylogenetic trees and three regular hybrids to prove that h_r does not satisfy the triangle inequality.

4.2 The rSPR distance is majorized by h

Given a rooted digraph (V, A) and any vertex v of V, $v \neq \rho$, let

$$end(v) = \{(u, v) : (u, v) \in A\}.$$

Lemma 4.2.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Let $\mathcal{H}$ be a hybrid phylogeny that displays $\mathcal{T}$ and $\mathcal{T}'$. There exists a hybrid phylogeny $\mathcal{H}'$ that displays both $\mathcal{T}$ and $\mathcal{T}'$ such that:*

(i) For all vertices v of $\mathcal{H}'$, $d^-(v) \leq 2$.

(ii) If $d^-(v) = 2$ for some vertex v of $\mathcal{H}'$ then one arc ending in v is used for the displaying of $\mathcal{T}$, the other is used for the displaying of $\mathcal{T}'$.

(iii) $h(\mathcal{H}') < h(\mathcal{H})$, unless, taking $\mathcal{H}' = \mathcal{H}$ already satisfies (i) and (ii).

Proof. For each vertex $v \in \mathcal{H}$, at most one arc of $end(v)$ is used for the displaying of $\mathcal{T}$, respectively $\mathcal{T}'$. Let v be a vertex of $V(\mathcal{H})$ such that either condition (i) or (ii) is not satisfied and suppose that at least one arc of $end(v)$ is used for the displaying. If (u, v) is an arc that is not used for the displaying, then delete the arc (u, v), adjoin a new leaf x via a new arc (u, x). If no arc of $end(v)$ is used for the displaying of $\mathcal{T}$ or $\mathcal{T}'$ then do the same operation for all the arcs in $end(v)$, except one arc. Note that the hybrid $\mathcal{H}'$ obtained by applying these operations displays both trees and $h(\mathcal{H}') < h(\mathcal{H})$. □

Lemma 4.2.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic trees. Let $\mathcal{H}$ be a hybrid that displays $\mathcal{T}$ and $\mathcal{T}'$ and has the minimum number of hybrid events. Then $\mathcal{H}$ satisfies the conditions (i) and (ii) in Lemma 4.2.1.*

Proof. Let $\mathcal{H}$ be a minimal hybrid that displays $\mathcal{T}$ and $\mathcal{T}'$. Suppose now that either condition (i) or condition (ii) is not satisfied. Then we can delete an arc ending in v, as in Lemma 4.2.1. We obtain a hybrid whose hybridization number is smaller, contradictory to the minimality of $\mathcal{H}$. □

Lemma 4.2.3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic trees. Let $\mathcal{H}$ be a hybrid that displays $\mathcal{T}$ and $\mathcal{T}'$, with $h(\mathcal{H}) \geq 1$. There exist a rooted binary phylogenetic tree $\mathcal{T}_1$ and a hybrid phylogeny $\mathcal{H}_1$ such that*

(i) $d_{rSPR}(\mathcal{T}, \mathcal{T}_1) = 1$,

(ii) $\mathcal{H}_1$ *displays* $\mathcal{T}_1$ *and* $\mathcal{T}'$, *and*

(iii) $h(\mathcal{H}_1) = h(\mathcal{H}) - 1$.

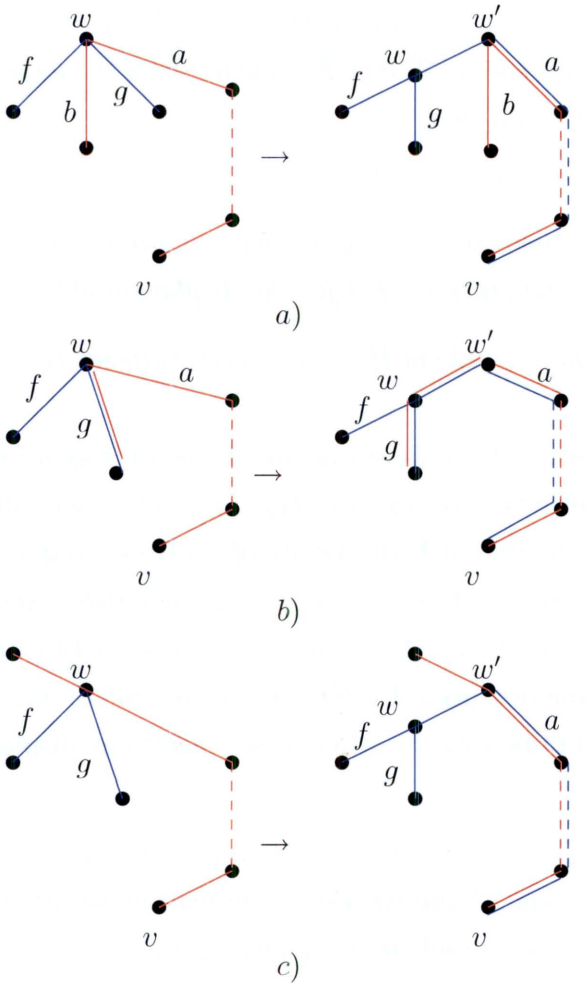


Figure 4.5: The popping operation in $\mathcal{H}'$. The 'blue' out–degree of w is two. In cases a) and b), the 'red' out–degree of w is two. In these cases w could be the root of $\mathcal{H}'$. In case c), the 'red' out–degree of w is one. It follows that the 'red' in–degree of w is 1.

Proof. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees and $\mathcal{H}$ a hybrid phylogeny that displays $\mathcal{T}$ and $\mathcal{T}'$. It follows from the definition of displaying that

$X \subseteq \mathcal{L}(\mathcal{H})$, and there exists a rooted subdigraph of $\mathcal{H}$ that is a refinement of $\mathcal{T}$, and a rooted subdigraph of $\mathcal{H}$ that is a refinement of $\mathcal{T}'$.

Colour by blue (respectively by red), the arcs of the rooted subdigraph of $\mathcal{H}$ that is a refinement of $\mathcal{T}$ (respectively $\mathcal{T}'$).

According to Lemma 4.2.1, we may assume that $d^-(v) \leq 2$ for any $v \in V(\mathcal{H})$ and if $d^-(v) = 2$, then one arc is used for the displaying of $\mathcal{T}$ and the other for the displaying of $\mathcal{T}'$. Also, at most two arcs starting in v are used for the displaying of $\mathcal{T}$, respectively of $\mathcal{T}'$. Let $v \in V(\mathcal{H})$ with $d^-(v) = 2$ (such v exists since $h(\mathcal{H}) \geq 1$),

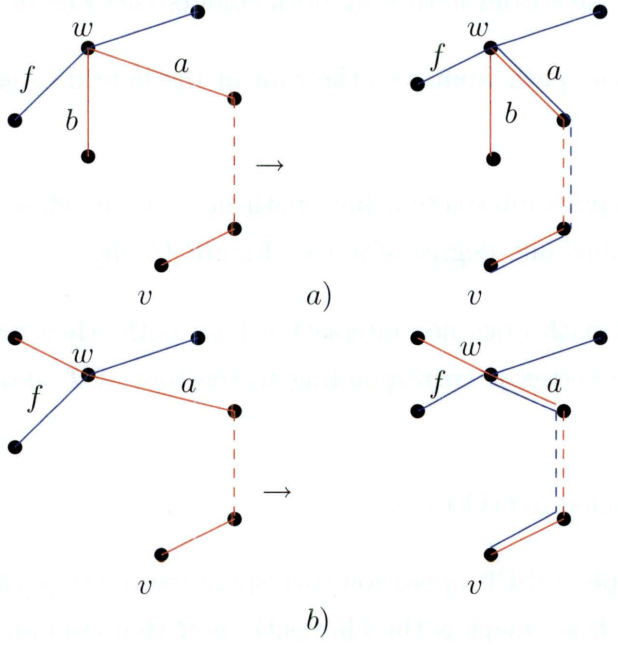


Figure 4.6: The 'blue' out–degree of w is one. It follows that the 'blue' in–degree of w is one.

and denote by (u_1, v) the (blue) arc used for the displaying of $\mathcal{T}$ and by (u_2, v) the (red) arc used for $\mathcal{T}'$. In the hybrid $\mathcal{H}$, delete the arc (u_1, v), and adjoin a new leaf x by a new arc (u_1, x). Denote the new hybrid by $\mathcal{H}'$. Follow up the red path until either:

1) it intersects a blue path (in a vertex) of $\mathcal{H}'$, or

2) it does not intersect a blue path; it reaches the vertex ρ corresponding to the

root of $\mathcal{T}'$ in $\mathcal{H}'$.

Case 1) Let w be the vertex in $\mathcal{H}'$ where the red path intersects the blue one. There are two possible cases:

Case 1a) Suppose the 'blue' out–degree of w is two (there exist two blue arcs starting in w). Then refine $\mathcal{H}'$ by applying the popping operation (see Figure 4.5). Colour (w', w) by blue and the path from w' to v by blue and red.

Case 1b) If the 'blue' out–degree of w is one (there is one blue arc starting with w), then colour the path from w to v by both colours (see Figure 4.6).

Case 2) Follow the path from ρ to the root of the hybrid. Again two cases are possible:

Case 2a) If this path intersects a blue path in w, then refine $\mathcal{H}'$ as in case 1), depending on the 'blue' out–degree of w (see Figure 4.7, b),c)).

Case 2b) If this path does not intersect a blue path, then colour by blue the paths from ρ to the vertex w corresponding to the root of $\mathcal{T}$ and to v (see Figure 4.7, a)).

Denote the obtained hybrid by $\mathcal{H}_1$.

In the tree $\mathcal{T}$ apply a rSPR operation corresponding to the paths in $\mathcal{H}$. Consider the arc (a, b) of $\mathcal{T}$ whose image is the blue path in $\mathcal{H}$ that contains (u_1, v). Cut the arc (a, b), and prune the subtree with the root b. If the 'blue' out-degree of w is 2, then w represents a vertex s of $\mathcal{T}$. Reattach the subtree up to the vertex s of $\mathcal{T}$ represented in $\mathcal{H}$ by w. If the 'blue' out–degree of w is 1, then the blue path represents an arc (c, d) of $\mathcal{T}$. Reattach the subtree up to d. □

Proposition 4.2.4. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Then*

$$d_{rSPR}(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}, \mathcal{T}').$$

Proof. If $h(\mathcal{T}, \mathcal{T}') = 0$ then $\mathcal{T} = \mathcal{T}'$, and therefore $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 0$.

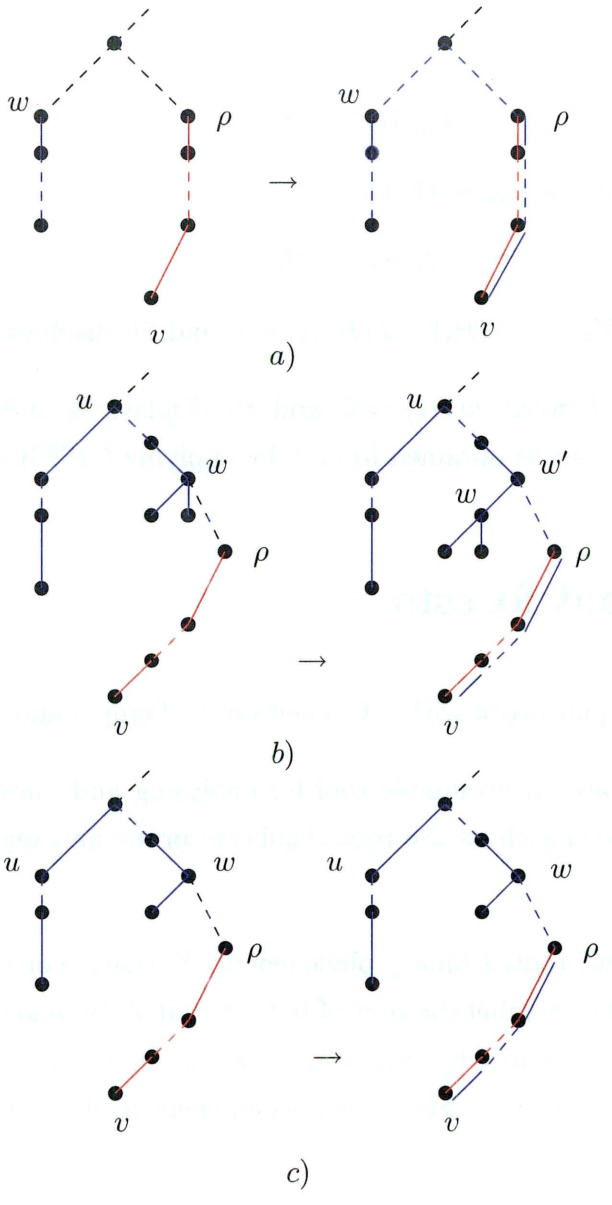


Figure 4.7:

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic trees such that $h(\mathcal{T}, \mathcal{T}') = k \geq 1$. Let $\mathcal{H}$ be a hybrid phylogeny such that $\mathcal{H}$ displays $\mathcal{T}$ and $\mathcal{T}'$ and $\mathcal{H}$ has the minimum number of hybrid events $h(\mathcal{H}) = h(\mathcal{T}, \mathcal{T}') = k$. We prove that $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by applying k rSPR operations. By succesively applying Lemma 4.2.3, one

can obtain a sequence of hybrids

$$\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_k$$

and a sequence of binary phylogenetic trees

$$\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$$

such that $d_{rSPR}(\mathcal{T}_{i-1}, \mathcal{T}_i) = 1$, $h(\mathcal{H}_i) = h(\mathcal{H}_{i-1}) - 1$, and $\mathcal{H}_i$ displays $\mathcal{T}_i$ and $\mathcal{T}'$.

In particular, after k steps, $h(\mathcal{H}_k) = 0$, and $\mathcal{H}_k$ displays $\mathcal{T}_k$ and $\mathcal{T}'$. It follows that $\mathcal{T}_k = \mathcal{T}'$, hence $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by applying k rSPR operations. □

4.3 Agreement forests

This section describes joint work with S.Grünewald, C. Semple and V. Moulton.

Agreement forests are an invaluable tool for analysing and understanding tree rearrangement operations such as the rooted subtree prune and regraft operation, as shown in [2, 11, 27].

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. For the purposes of the following definitions, we label the root of both $\mathcal{T}$ and $\mathcal{T}'$ by ρ and regard it as a vertex at the end of a pendant edge adjoined to the original root. Furthermore, in addition to the elements of X, we also view ρ as an element of the label set of both $\mathcal{T}$ and $\mathcal{T}'$.

An **agreement forest** for $\mathcal{T}$ and $\mathcal{T}'$ is a collection $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$, where $\mathcal{T}_\rho$ is a rooted tree and $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ are rooted binary phylogenetic trees with label sets $\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots \mathcal{L}_k$ such that the following properties hold:

(i) The label sets $\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots \mathcal{L}_k$ partition $X \cup \{\rho\}$ and, in particular, $\rho \in \mathcal{L}_\rho$.

(ii) For all $i \in \{\rho, 1, 2, \ldots, k\}$, $\mathcal{T}_i \cong \mathcal{T}|\mathcal{L}_i \cong \mathcal{T}'|\mathcal{L}_i$.

(iii) The trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ and $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are vertex disjoint rooted subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

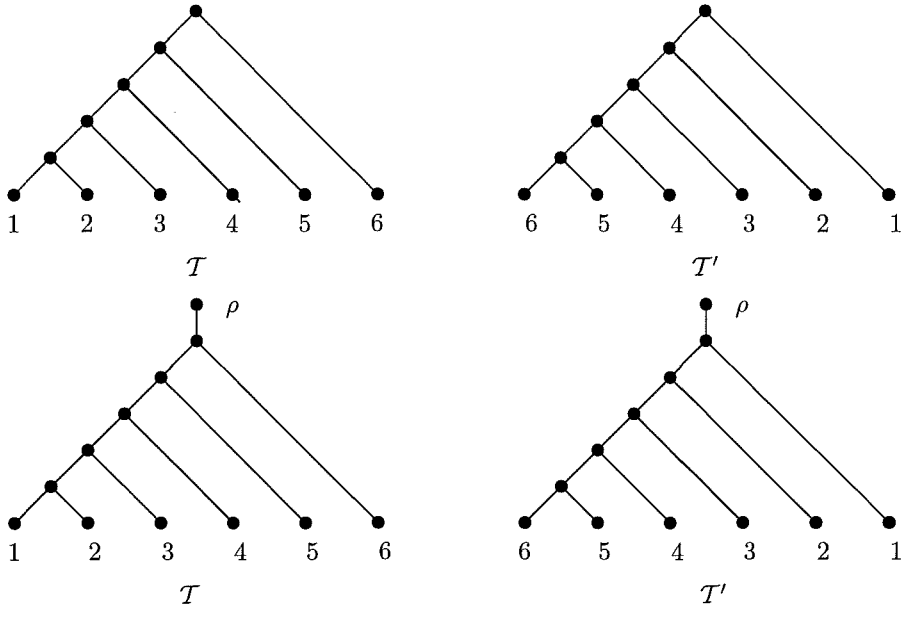


Figure 4.8: Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ without (above) and with (below) their roots labelled.

A **maximum agreement forest** for $\mathcal{T}$ and $\mathcal{T}'$ is an agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ for which $|\mathcal{F}|$ is minimal. Let

$$m(\mathcal{T}, \mathcal{T}') = \min\{|\mathcal{F}| - 1 : \mathcal{F} \text{ is an agreement forest for } \mathcal{T} \text{ and } \mathcal{T}'\}.$$

This definition of an agreement forest has been introduced in [11] and it has been used to prove that

$$d_{rSPR}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}').$$

We will show later in this section that a similar result can be obtained for $h(\mathcal{T}, \mathcal{T}')$ and a particular type of agreement forest. We now introduce this particular type of agreement forest. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees and let $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{G}_\mathcal{F}$ be the directed graph whose vertex set is $\mathcal{F}$ and for which $(\mathcal{T}_i, \mathcal{T}_j)$ is an arc precisely if one of the following conditions holds:

(A_1) the root of $\mathcal{T}(\mathcal{L}(\mathcal{T}_i))$ is an ancestor of the root of $\mathcal{T}(\mathcal{L}(\mathcal{T}_j))$, or

(A_2) the root of $\mathcal{T}'(\mathcal{L}(\mathcal{T}_i))$ is an ancestor of the root of $\mathcal{T}'(\mathcal{L}(\mathcal{T}_j))$.

We call $\mathcal{F}$ a **good agreement forest** if $\mathcal{G}_{\mathcal{F}}$ does not contain a directed cycle. A **maximum good agreement forest** for $\mathcal{T}$ and $\mathcal{T}'$ is a good agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ for which $|\mathcal{F}|$ is minimal. We denote by

$$m_g(\mathcal{T}, \mathcal{T}') = \min\{|\mathcal{F}| - 1 : \mathcal{F} \text{ is a good agreement forest for } \mathcal{T} \text{ and } \mathcal{T}'\}.$$

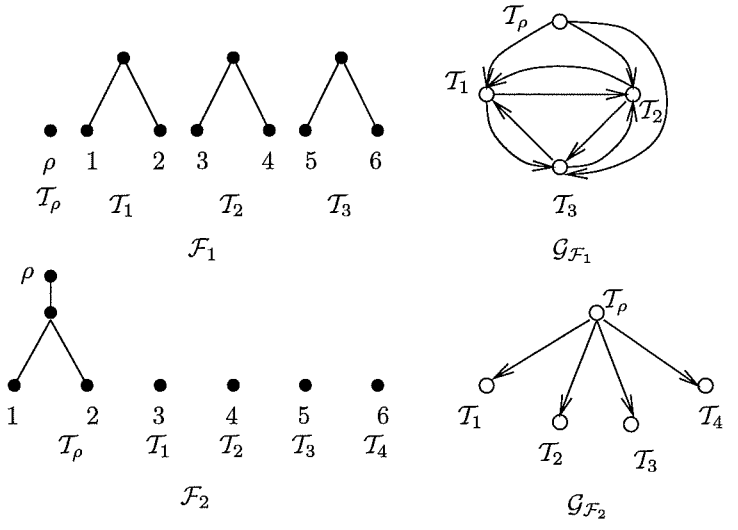


Figure 4.9: Two agreement forests for the trees $\mathcal{T}$ and $\mathcal{T}'$ shown in Figure 4.8. The forest $\mathcal{F}_1$ is a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ but is not a good agreement forest as the directed graph $\mathcal{G}_{\mathcal{F}_1}$ has directed cycles. The forest $\mathcal{F}_2$ is a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.

Note that in Figure 4.9, $\mathcal{F}_2$ is a maximum good agreement forest for the trees $\mathcal{T}$ and $\mathcal{T}'$ but $\mathcal{F}_1$ is not a good agreement forest for these two trees. Clearly, as every good agreement forest is an agreement forest,

$$m(\mathcal{T}, \mathcal{T}') \leq m_g(\mathcal{T}, \mathcal{T}').$$

The example in Figure 4.9 shows that this inequality may be strict.

Lemma 4.3.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees and let $\mathcal{F}$ be a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{T}_\rho \in \mathcal{F}$ be the rooted tree whose label set contains ρ. Then $\mathcal{L}(\mathcal{T}_\rho) \cap X$ is nonempty.*

Proof. Suppose that $\mathcal{L}(\mathcal{T}_\rho) \cap X = \emptyset$, that is $\mathcal{T}_\rho = \{\rho\}$. Since $\mathcal{F}$ is a good agreement forest, there exists a vertex $\mathcal{T}_0$ of $\mathcal{G}_\mathcal{F} \backslash \mathcal{T}_\rho$ with in–degree zero. It follows that the forest obtained from $\mathcal{F}$ by adding ρ at the end of a pendant edge adjoined to the root of $\mathcal{T}_0$ is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with one less component, a contradiction with the minimality of $|\mathcal{F}|$. □

Theorem 4.3.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Then*

$$h(\mathcal{T}, \mathcal{T}') = m_g(\mathcal{T}, \mathcal{T}').$$

Proof. We prove first that $h(\mathcal{T}, \mathcal{T}') \geq m_g(\mathcal{T}, \mathcal{T}')$. Let $\mathcal{H}$ be a hybrid that displays $\mathcal{T}$ and $\mathcal{T}'$ and has a minimum number of hybrid events. By Lemma 4.2.2 we may assume that, for each vertex of $\mathcal{H}$, the number of incoming arcs is at most two. Let $\mathcal{F}$ be the forest obtained from $\mathcal{H}$ by deleting, for each hybridization vertex v of $\mathcal{H}$, the two incoming arcs and then supressing any resulting degree-two vertex. We show by induction on $h(\mathcal{H})$ that $\mathcal{F}$ is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with $h(\mathcal{H}) + 1$ components, and therefore $h(\mathcal{T}, \mathcal{T}') \geq m_g(\mathcal{T}, \mathcal{T}')$.

If $h(\mathcal{H}) = 0$, then $\mathcal{T}$ and $\mathcal{T}'$ are—up to isomorphism—identical, so the result holds. Now let $h(\mathcal{H}) = n$ and assume that the result holds for all pairs of rooted binary phylogenetic X–trees for which $h(\mathcal{H})$ is at most $n - 1$, where $n \geq 1$. Let v be a hybridization vertex of $\mathcal{H}$ such that $\mathcal{H}|c(v)$ is a rooted binary phylogenetic tree T_v on $c(v)$. Such vertex must exist since $h(\mathcal{H}) \geq 1$. Then by Lemma 4.2.2, one of the arcs coming in to v, e_1 say, is used by $\mathcal{H}$ to display $\mathcal{T}_1$ and the other arc coming to v, e_2 say, is used by $\mathcal{H}$ to display $\mathcal{T}_2$.

Consider now a hybrid $\mathcal{H}'$ on X and two rooted binary phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_1'$ defined in the following way:

- Viewing the root ρ of $\mathcal{H}$ as a vertex at the end of a pendant edge adjoined to the original root, $\mathcal{H}'$ is obtained from $\mathcal{H}$ by deleting e_1 and e_2, and then adjoining the root of $\mathcal{T}_v$ to ρ.

- Viewing the root of $\mathcal{T}$ as a vertex ρ at the end of a pendant edge adjoined to the original root, $\mathcal{T}_1$ is obtained from $\mathcal{T}$ by pruning the rooted subtree $\mathcal{T}_v$

and adjoining the root of this subtree to ρ with a new edge. Similarly, $\mathcal{T}_1'$ is obtained from $\mathcal{T}'$.

Clearly, $h(\mathcal{H}') = h(\mathcal{H}) - 1 = n - 1$. Moreover, as $\mathcal{H}$ displays both $\mathcal{T}$ and $\mathcal{T}'$, it follows from the above construction that $\mathcal{H}'$ displays both $\mathcal{T}_1$ and $\mathcal{T}_1'$. Therefore, by the induction assumption, the forest $\mathcal{F}'$ obtained from $\mathcal{H}'$ by deleting, for each hybridization vertex, the two incoming arcs and then supressing any resulting degree–two vertex is a good agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$ with n components.

Now let us observe that at most one tree in $\mathcal{F}'$ has the property that its label set contains elements of both $c(v)$ and $X - c(v)$, in which case this tree is $\mathcal{T}_\rho$ (the component of $\mathcal{F}'$ that contains ρ). By the maximality of $\mathcal{F}'$, it follows that the label set of $\mathcal{T}_\rho$ contains $c(v)$.

Let $\mathcal{F}$ be the forest obtained from $\mathcal{F}'$ by deleting the edge joining ρ to the root of $\mathcal{T}_v$. Since $\mathcal{F}'$ is a good agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$, $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, as both $\mathcal{T}$ and $\mathcal{T}'$ contain $\mathcal{T}_v$ as a rooted subtree, it follows that $\mathcal{F}$ is also a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Since $\mathcal{F}$ has $n+1$ components, we deduce that $h(\mathcal{T}, \mathcal{T}') \geq m_g(\mathcal{T}, \mathcal{T}')$.

We next show that $m_g(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}, \mathcal{T}')$. The proof is by induction on $m_g(\mathcal{T}, \mathcal{T}')$. If $m_g(\mathcal{T}, \mathcal{T}') = 0$, then, up to isomorphism, $\mathcal{T}$ and $\mathcal{T}'$ are identical, so therefore $m_g(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}, \mathcal{T}')$. Now let $m_g(\mathcal{T}, \mathcal{T}') = k$ and assume that the result holds for all pairs of rooted binary phylogenetic X–trees for which the minimum number of components over all good agreement forests is at most k.

Let $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Since $\mathcal{F}$ is good, $\mathcal{G}_\mathcal{F}$ has no directed cycles and so there exists a vertex of $\mathcal{G}_\mathcal{F}$ whose out–degree is zero. Without loss of generality, we may assume that this vertex is $\mathcal{T}_k$. It follows that $\mathcal{T}_k$ is a rooted subtree of both $\mathcal{T}$ and $\mathcal{T}'$. Let $X_k = X - \mathcal{L}(\mathcal{T}_k)$ and let $\mathcal{F}_k = \mathcal{F} - \{\mathcal{T}_k\}$. Then $\mathcal{F}_k$ is a good agreement forest for $\mathcal{T}|X_k$ and $\mathcal{T}'|X_k$. Since $|\mathcal{F}_k| < |\mathcal{F}|$, it follows by the induction assumption that there is a hybrid $\mathcal{H}_k$ on X that displays both $\mathcal{T}|X_k$ and $\mathcal{T}'|X_k$, and has the property that $h(\mathcal{H}_k) \leq k - 1$.

Since $\mathcal{H}_k$ displays $\mathcal{T}|X_k$ and since $\mathcal{T}_k$ is a rooted subtree of $\mathcal{T}$, there exists a

hybrid that can be obtained from $\mathcal{H}_k$ by adjoining $\mathcal{T}_k$ via a new edge e that connects the root of $\mathcal{T}_k$ and a new vertex that subdivides an edge of $\mathcal{H}_k$. Similarly, there is a hybrid that displays $\mathcal{T}'|X_k$ and that can be obtained from $\mathcal{H}_k$ by adjoining $\mathcal{T}_k$ using a new edge e'. Now let $\mathcal{H}$ be the hybrid obtained from $\mathcal{H}_k$ by adjoining $\mathcal{T}_k$ using exactly edges e and e'. Clearly, $\mathcal{H}_k$ displays both $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, since $\mathcal{T}_k$ is a rooted binary phylogenetic tree and the vertex of $\mathcal{H}$ corresponding to the root of $\mathcal{T}_k$ has in–degree two, $h(\mathcal{H}) \leq k$. Hence $h(\mathcal{T}, \mathcal{T}') \leq k = m_g(\mathcal{T}, \mathcal{T}')$. This completes the proof of the theorem. □

4.4 Bounds for $h(\mathcal{T}, \mathcal{T}')$

Note that by Theorem 4.3.2 and previous observations, we obtain an alternative proof for

$$d_{rSPR}(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}, \mathcal{T}').$$

The following proposition shows that the lower bound for h is sharp.

Proposition 4.4.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Then $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 1$ if and only if $h(\mathcal{T}, \mathcal{T}') = 1$. Moreover, if $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 1$, a regular hybrid that displays both trees can be easily obtained.*

Proof. Assume that $h(\mathcal{T}, \mathcal{T}') = 1$. From Proposition 4.2.4 and the properties of d_{rSPR} and h it follows that $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 1$.

Now, let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees such that $\mathcal{T}'$ is obtained from $\mathcal{T}$ by a single $rSPR$ operation $\theta_{a,c}$, which prunes the subtree with the root a and regrafts it up to the vertex c. We will describe now how we can construct a regular hybrid that displays both trees. Consider the following sequence of operations:

(1) If c is not the root of $\mathcal{T}$ then there exists a vertex b of $\mathcal{T}$ that is a direct ancestor of c.

(a) Subdivide the arc (b, c) by a new vertex d and add a new arc (d, a).

(b) In the case of a 'down' rSPR operation, subdivide the arc (u, a) and adjoin a new leaf x to the subdividing vertex.

(c) In the case of an 'up' rSPR operation, subdivide the arc (d, a) and adjoin a new leaf x to the subdividing vertex.

(2) If $c = \rho_T$ then add a new vertex b, a new arc (b, c) and a new arc (b, a). Subdivide the arc (b, a) and adjoin a new leaf x to the subdividing vertex.

The hybrid $\mathcal{H}$ obtained by performing the previous set of operations is regular, displays $\mathcal{T}$ and $\mathcal{T}'$, and $h(\mathcal{H}) = 1$. For each case the construction is illustrated in Figure 4.10. □

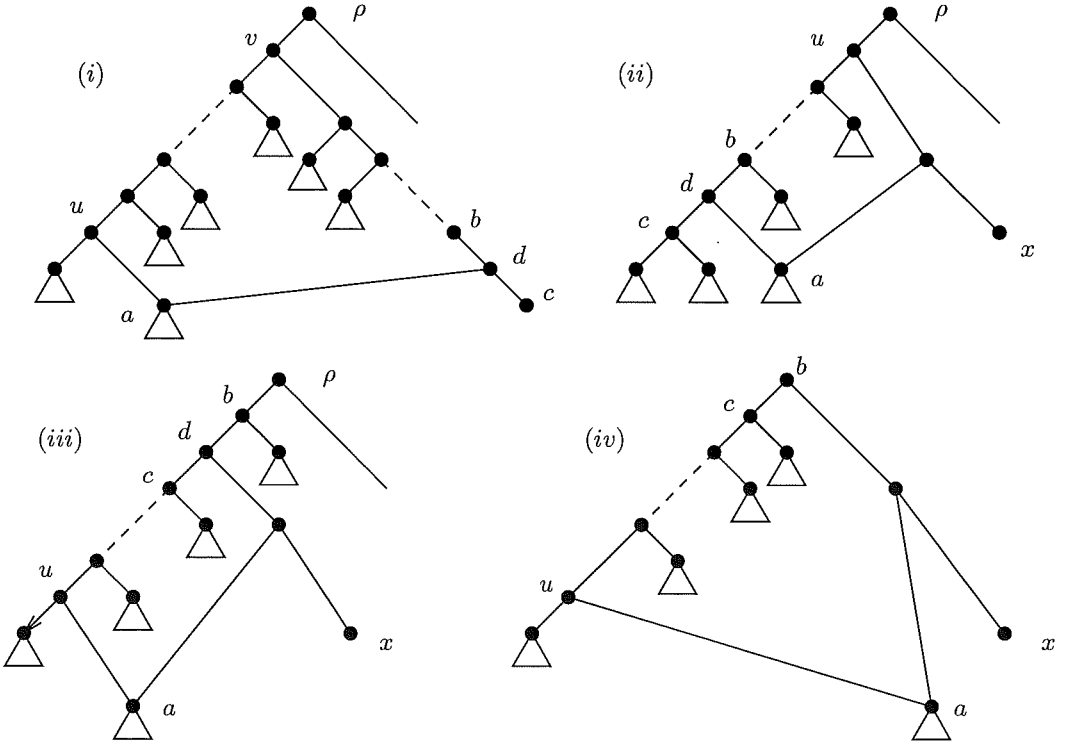


Figure 4.10: Each figure (i)–(iv) shows a regular hybrid that displays two trees $\mathcal{T}$ and $\mathcal{T}'$ with $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 1$.

In a biological interpretation, the new leaf x is a species not mentioned by any of the input trees (one that may be now extinct) that may have been involved in the hybridization in the past (see [36]).

In the next proposition we establish an upper bound for $h(\mathcal{T}, \mathcal{T}')$.

Proposition 4.4.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Then*

$$h(\mathcal{T}, \mathcal{T}') \leq |X| - 2.$$

Proof. Let $n = |X|$. To prove that $h(\mathcal{T}, \mathcal{T}') \leq |X| - 2$, we will construct a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with $n - 1$ components. The result then follows by applying Proposition 4.3.2.

To construct a good agreement forest, let us observe that for each tree, the root partitions the set of leaves in two disjoint proper subsets, say A_1, A_2 for $\mathcal{T}$ and A_1', A_2' for $\mathcal{T}'$. Without loss of generality, we may assume that $A_1 \cap A_1' \neq \emptyset$ and $A_2 \cap A_2' \neq \emptyset$. Let $x \in A_1 \cap A_1'$ and $y \in A_2 \cap A_2'$. Denote by $\mathcal{T}_\rho$ the rooted binary tree with the root ρ and the leaves x and y, and for all $i \in \{1, 2, \ldots, n-2\}$, let $\mathcal{T}_i$ be the rooted phylogenetic tree consisting of a single vertex labelled by l_i, $l_i \in X - \{x, y\}$. Then the set $\{\mathcal{T}_\rho, \mathcal{T}_1, \ldots, \mathcal{T}_{n-2}\}$ is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with $n - 1$ components. □

The next proposition shows that the upper bound for h is sharp.

Proposition 4.4.3. *For all $n \geq 2$ there exist two rooted binary X–trees $\mathcal{T}$ and $\mathcal{T}'$ with $|X| = n$ and $h(\mathcal{T}, \mathcal{T}') = n - 2$.*

Proof. Let $X = \{x_1, x_2, \ldots, x_n\}$ and let $\mathcal{T}$ and $\mathcal{T}'$ be the two rooted binary phylogenetic X–trees shown in Figure 4.11. We will prove that $m_g(\mathcal{T}, \mathcal{T}') = n - 2$. Let $\mathcal{F}$ be a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ and let $\mathcal{T}_\rho$ be the tree in $\mathcal{F}$ that has ρ as a vertex label.

First we make the following observations:

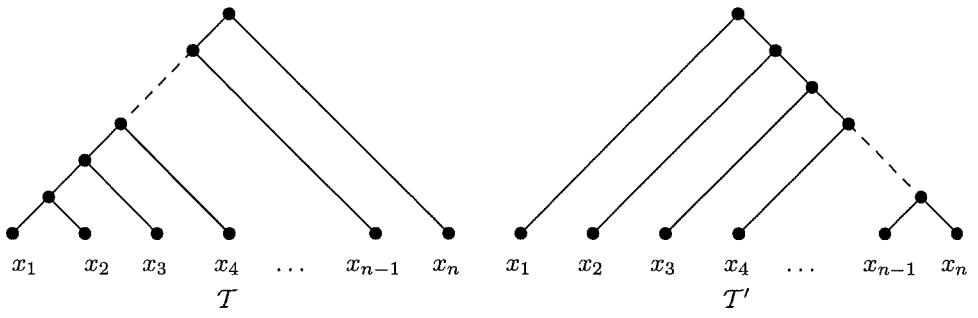


Figure 4.11: Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ with n leaves such that $h(\mathcal{T}, \mathcal{T}') = n - 2$.

(1) Each tree in $\mathcal{F}$ has a label set that contains at most two elements of X, for otherwise $\mathcal{F}$ is not an agreement forest (condition (ii) in the definition is not satisfied).

(2) There exists at most one tree in $\mathcal{F}$ with the property that its label set contains two elements from X. To see this, suppose that there are two such trees, $\mathcal{T}_i$ and $\mathcal{T}_j$, in $\mathcal{F}$. Since the trees in $\{\mathcal{T}(\mathcal{L}(\mathcal{T}_i)), \mathcal{T}(\mathcal{L}(\mathcal{T}_j))\}$ and $\{\mathcal{T}'(\mathcal{L}(\mathcal{T}_i)), \mathcal{T}'(\mathcal{L}(\mathcal{T}_j))\}$ are vertex disjoint subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively, it follows that neither $\mathcal{L}(\mathcal{T}_i)$ nor $\mathcal{L}(\mathcal{T}_j)$ contains ρ. Therefore $\mathcal{L}(\mathcal{T}_i) = \{i_1, i_2\}$, $\mathcal{L}(\mathcal{T}_j) = \{j_1, j_2\}$ and $i_1 < i_2 < j_1 < j_2$. But then by considering the vertices in $\mathcal{T}$ and $\mathcal{T}'$ corresponding to the roots of $\mathcal{T}_i$ and $\mathcal{T}_j$ we obtain that $\mathcal{G}_{\mathcal{F}}$ contains a directed cycle, a contradiction.

(3) By Proposition 4.3.1, the label set of $\mathcal{T}_\rho$ contains at least one element of X.

It follows that either the forest contains one tree with three vertices (one of them being ρ) or else it contains exactly two trees with two vertices. In the former case, there are $n - 2$ isolated leaves, hence $m_g(\mathcal{T}, \mathcal{T}') = n - 2$. In the latter one, there are $n - 1$ components: one containing ρ and a leaf, one containing two leaves, and $n - 3$ isolated leaves. Consequently, $m_g(\mathcal{T}, \mathcal{T}') = n - 2$. □

4.5 When does d_{rSPR} equal h?

Though the maximal difference between h and d_{rSPR} can be large for large n (as we will show in Section 4.6), they are equal for small enough trees.

Proposition 4.5.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. If $1 \leq n = |X| \leq 5$, then*

$$d_{rSPR}(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}, \mathcal{T}').$$

Proof. The statement clearly holds for $n \in \{1, 2\}$.

If $n = 3$, at most one $rSPR$ operation can be performed, so according to Proposition 4.4.1, we have $d_{rSPR}(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}, \mathcal{T}')$. For $n = 4$, at most two $rSPR$ operations can be applied. If $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 2$, then $d_{rSPR}(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}, \mathcal{T}') = 4-2 = 2$, so we obtain the equality.

If $n = 5$, at most three $rSPR$ operations can be performed. If $d_{rSPR}(\mathcal{T}, \mathcal{T}') \in \{1, 3\}$, then from Proposition 4.4.1 and respectively Corollary 4.2.4, it follows that $h(\mathcal{T}, \mathcal{T}')$ is equal to 1, respectively to 3.

We will prove that if $h(\mathcal{T}, \mathcal{T}') = 3$ then $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 3$. Assume that $h(\mathcal{T}, \mathcal{T}') = 3$ and $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 2$. Then there exists a maximum agreement forest $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2\}$ for $\mathcal{T}$ and $\mathcal{T}'$ that is not a good agreement forest. Only the following cases are possible:

Case 1) The tree $\mathcal{T}_\rho$ is the rooted phylogenetic tree consisting of a single vertex labelled by ρ, and $\mathcal{T}_1$ and $\mathcal{T}_2$ are rooted phylogenetic trees with two, respectively three leaves labelled by elements of X.

By adding the root ρ to $\mathcal{T}_2$ and considering the leaves of $\mathcal{T}_1$ as isolated vertices, we obtain a good agreement forest with only two components not containing the root. It follows that $h(\mathcal{T}, \mathcal{T}') = 2$, a contradiction.

Case 2) The tree $\mathcal{T}_\rho$ is a rooted X'–tree with one leaf labelled by an element of X and the root labelled by ρ. Both $\mathcal{T}_1$ and $\mathcal{T}_2$ are rooted phylogenetic trees with two leaves labelled by elements of X. If we adjoin $\mathcal{T}_1$ to $\mathcal{T}_\rho$ and take the leaves of

$\mathcal{T}_2$ as isolated vertices, we obtain a good agreement forest, and again it follows that $h(\mathcal{T}, \mathcal{T}') = 2$, a contradiction. □

Note that $h(\mathcal{T}, \mathcal{T}') = d_{rSPR}(\mathcal{T}, \mathcal{T}')$ whenever there exists a maximum agreement forest that is also a good agreement forest. In Chapter 5 we will discuss this in detail and we will show how we can construct a minimal regular hybrid $\mathcal{H}$ that displays $\mathcal{T}$ and $\mathcal{T}'$ starting from an appropriate sequence of rSPR operations.

4.6 How large can $h(\mathcal{T}, \mathcal{T}') - d_{rSPR}(\mathcal{T}, \mathcal{T}')$ be?

In this section, we will construct pairs of rooted binary phylogenetic X–trees $(\mathcal{T}, \mathcal{T}')$ for which the difference between $h(\mathcal{T}, \mathcal{T}')$ and $d_{rSPR}(\mathcal{T}, \mathcal{T}')$ is large.

Proposition 4.6.1. *For any $n \geq 4$ there exist two rooted binary X–trees $\mathcal{T}$ and $\mathcal{T}'$ with $|X| = n$ and*

$$d_{rSPR}(\mathcal{T}, \mathcal{T}') = \lfloor \frac{n+1}{2} \rfloor$$

and

$$h(\mathcal{T}, \mathcal{T}') = n - 2.$$

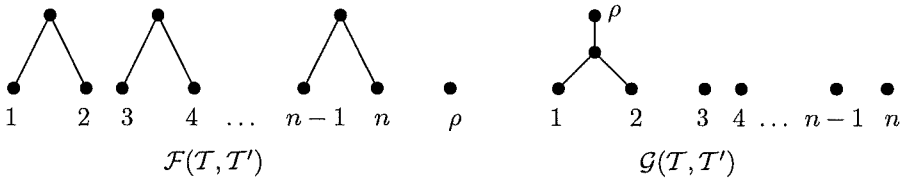
Proof. Let $\mathcal{T}$ and $\mathcal{T}'$ be the two trees in Figure 4.11. We have proved (Proposition 4.4.3) that $h(\mathcal{T}, \mathcal{T}') = n - 2$.

We show now that $d_{rSPR}(\mathcal{T}, \mathcal{T}') = \lfloor \frac{n+1}{2} \rfloor$. Suppose that n is even. First, let us observe that an agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ can be obtained by taking each adjacent pair $2i-1$, $2i$ ($i \in \{1, 2, \ldots, n/2\}$) and ρ to be the label sets of the $n/2+1$ components of $\mathcal{F}$. Thus, $d_{rSPR}(\mathcal{T}, \mathcal{T}') \leq n/2$.

In order to prove that $d_{rSPR}(\mathcal{T}, \mathcal{T}') \geq n/2$, we assume that there exists a maximum agreement forest $\mathcal{F}'$ for $\mathcal{T}$ and $\mathcal{T}'$ with less than $n/2+1$ components. If the component $\mathcal{T}_\rho$ of $\mathcal{F}'$ contains also a leaf $i \in X$ for some $i \in \{1, 2, \ldots, n\}$ then all the other leaves have to be isolated vertices of $\mathcal{F}'$. Since each tree of an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ contains at most two vertices labelled by elements of X, it follows that $|\mathcal{F}'| \in \{n-1, n\}$, so $|\mathcal{F}'| \geq n/2+1$, a contradiction. Therefore, ρ labels an isolated

vertex of $\mathcal{F}'$. Since $|\mathcal{F}'| < n/2 + 1$ it follows that there exists a component of $\mathcal{F}'$ with at least three leaves labelled by elements of X, a contradiction, for in this case $\mathcal{F}'$ is not an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.

The proof in the case where n is odd is similar. □



$\mathcal{F}(\mathcal{T},\mathcal{T}')$ $\mathcal{G}(\mathcal{T},\mathcal{T}')$

Figure 4.12: A maximum agreement forest $\mathcal{F}(\mathcal{T},\mathcal{T}')$, and a maximum good agreement forest $\mathcal{G}(\mathcal{T},\mathcal{T}')$ for the trees in Figure 4.11.

Proposition 4.6.2. *For any $n \geq 4$ there exist two rooted binary X–trees $\mathcal{T}$ and $\mathcal{T}'$ with $|X| = n$ and such that*

$$d_{rSPR}(\mathcal{T},\mathcal{T}') = 2$$

and

$$h(\mathcal{T},\mathcal{T}') = \lfloor n/2 \rfloor.$$

Proof. Let $n \geq 4$ and $X = \{1, 2, \ldots, n\}$ and consider the two trees shown in Figure 4.13, where $k = \lfloor n/2 \rfloor$. The tree $\mathcal{T}'$ is obtained from $\mathcal{T}$ by applying two $rSPR$ operations. In the first operation, the subtree with the set of leaves $X_1 = \{1, 2, \ldots, k\}$ is pruned and reattached up to the root. For the second operation, the subtree with the leaves $X_2 = \{k+1, k+2, \ldots, n\}$ is pruned and reattached to the arc ending in 1. A maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is shown in Figure 4.14.

We will show that a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ should have at least k components that do not contain the root.

First, let us observe that $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \ldots \mathcal{T}_k\}$, where $\mathcal{T}_\rho$ is the tree obtained by adding ρ to $\mathcal{T}|X_2$, and $\mathcal{T}_i$ is the leaf labelled by i, $i \in X_1$, is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.

Then we make the following observations:

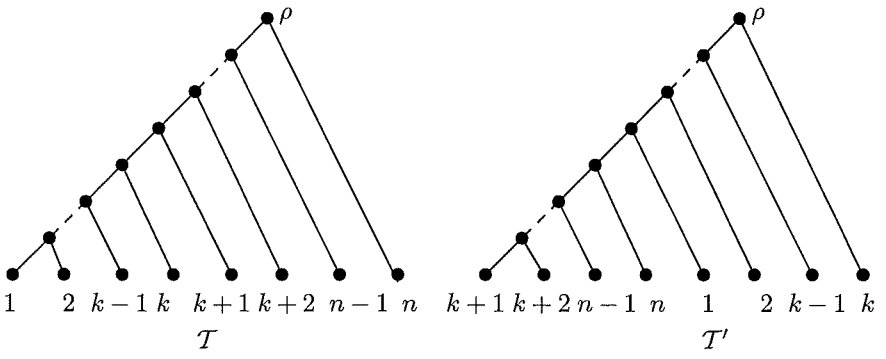


Figure 4.13: Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ such that $h(\mathcal{T}, \mathcal{T}') = \lfloor n/2 \rfloor$.

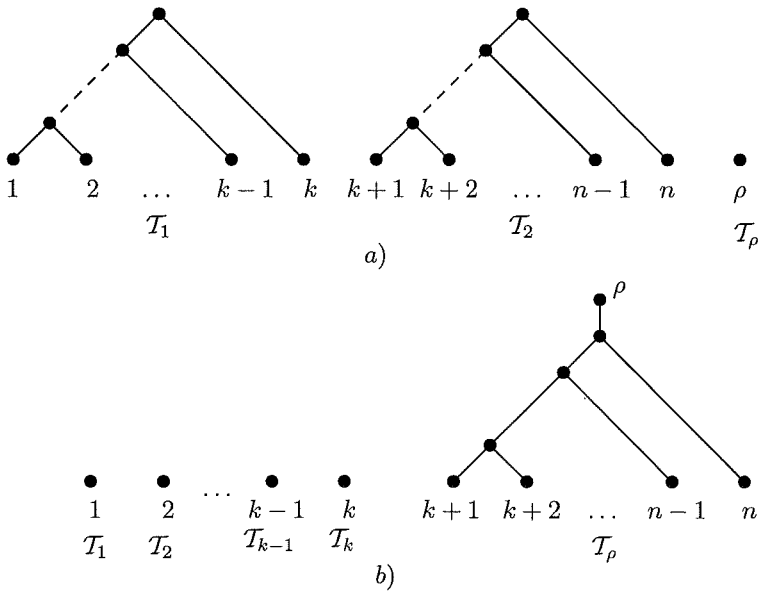


Figure 4.14: a) A maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. b) A maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.

(i) In an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, any tree with more than three leaves should have all the leaves either in X_1 or in X_2.

(ii) If an agreement forest contains one tree with two leaves i and j, $i \in X_1$ and $j \in X_2$, then all the other components with elements from X should be isolated vertices (leaves). (The root ρ is added to the component with the leaves i, j.)

(iii) A good agreement forest cannot contain one tree with more than two leaves from X_1 and one tree with more than two leaves from X_2.

It follows from (i)–(iii) that we can construct a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ in the following way: let $\mathcal{T}_\rho$ be the tree obtained by adding ρ to $\mathcal{T}|X_2$ and the other components as isolated vertices (leaves). Hence $\mathcal{F}$ is a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ and therefore $h(\mathcal{T}, \mathcal{T}') = k$. □

Note that the result in Proposition 4.11 slightly improves the lower bound obtained in [51] for the diameter of the space of rooted binary phylogenetic trees, measured using d_{rSPR}.

We can use Proposition 4.4.1 and Proposition 4.6.2 to show that h and h_r^+ do not satisfy the triangle inequality.

Proposition 4.6.3. *There exist three rooted binary phylogenetic trees $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ such that*

$$h(\mathcal{T}_1, \mathcal{T}_3) > h(\mathcal{T}_1, \mathcal{T}_2) + h(\mathcal{T}_2, \mathcal{T}_3).$$

Proof. Consider the rooted binary phylogenetic trees shown in Figure 4.15. Note that $\mathcal{T}_1$ and $\mathcal{T}_3$ are the phylogenetic trees from Proposition 4.6.2 in the particular case when $n = 6$. [1] Then $d_{rSPR}(\mathcal{T}_1, \mathcal{T}_2) = d_{rSPR}(\mathcal{T}_2, \mathcal{T}_3) = 1$ and therefore $h(\mathcal{T}_1, \mathcal{T}_2) = h(\mathcal{T}_2, \mathcal{T}_3) = 1$. On the other hand, $h(\mathcal{T}_1, \mathcal{T}_3) = 3$. □

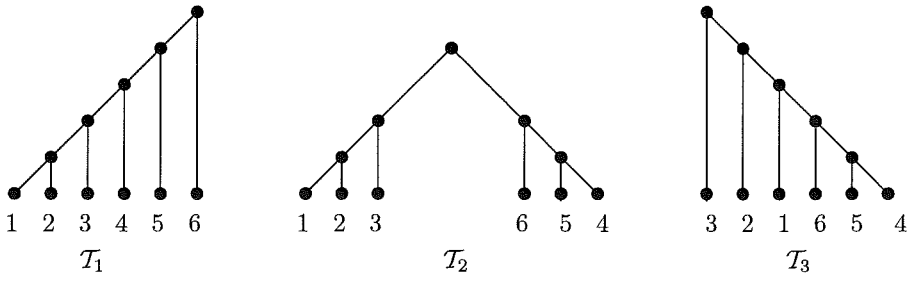


Figure 4.15: Three rooted binary phylogenetic trees to prove that h does not satisfy the triangle inequality.

[1] Note that at least six leaves are needed to construct such an example.

4.7 Some remarks on $h_r(\mathcal{T}_1, \mathcal{T}_2)$

This section is joint work with S. Grünewald, K.Huber and V. Moulton [6].

For two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$,

$$h_r^+(\mathcal{T}_1, \mathcal{T}_2) \leq h_r(\mathcal{T}_1, \mathcal{T}_2) \leq h(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]).$$

Is it possible to compute $h_r(\mathcal{T}_1, \mathcal{T}_2)$ starting from $h(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2])$? More precisely, can we construct a regular hybrid that displays two trees with a minimum number of hybrid events starting from the cluster union hybrid ? Or, equivalently, starting from the union of clusters of the two trees, can we define an operation on the collection of clusters such that, by successively applying this operation, we obtain a set of clusters corresponding to a minimal regular hybrid that displays the two trees?

Let us first introduce some definitions and notations. Let $\mathcal{C}$ be a collection of clusters on X. If $(A, B) \in \mathcal{C} \times \mathcal{C}$ such that $B \subsetneq A$ we say that A is a **parent** of B, and B is a **child** of A. Denote by

$$L_{\mathcal{C}} = \{(P, C) \in \mathcal{C} \times \mathcal{C} | C \text{ is a child of } P\}.$$

Let $\mathcal{C}$ and $\mathcal{C}'$ be two cluster systems on X. We say that $\mathcal{C}'$ **properly displays** $\mathcal{C}$ if the following conditions hold:

(P1) For any $C \in \mathcal{C}$ there exists $Q_C \in \mathcal{C}'$ such that $C \subseteq Q_C$ and $Q_C \subseteq Q$ for all $Q \in \mathcal{C}'$ such that $C \subseteq Q$. (Q_C is well-defined for any $C \in \mathcal{C}$.)

(P2) $Q_C = Q_{C'} \Rightarrow C = C'$.

(P3) Let $C_1, C_2, P_1, P_2 \in \mathcal{C}$, $C_1 \neq C_2$ such that for every $i \in \{1, 2\}$, C_i is a child of P_i. Then, for $Q \in \mathcal{C}'$, we have

$$Q_{C_1} \cup Q_{C_2} \subseteq Q \subseteq Q_{P_1} \cap Q_{P_2} \Rightarrow Q \in \{Q_{C_1}, Q_{P_1}\} \cap \{Q_{C_2}, Q_{P_2}\}.$$

Note that from (P1)-(P3) it follows that $Q_C = Q_{C'}$ holds if and only if $C = C'$ holds and that $Q_C \subsetneq Q_{C'}$ holds whenever $C \subsetneq C'$ holds.

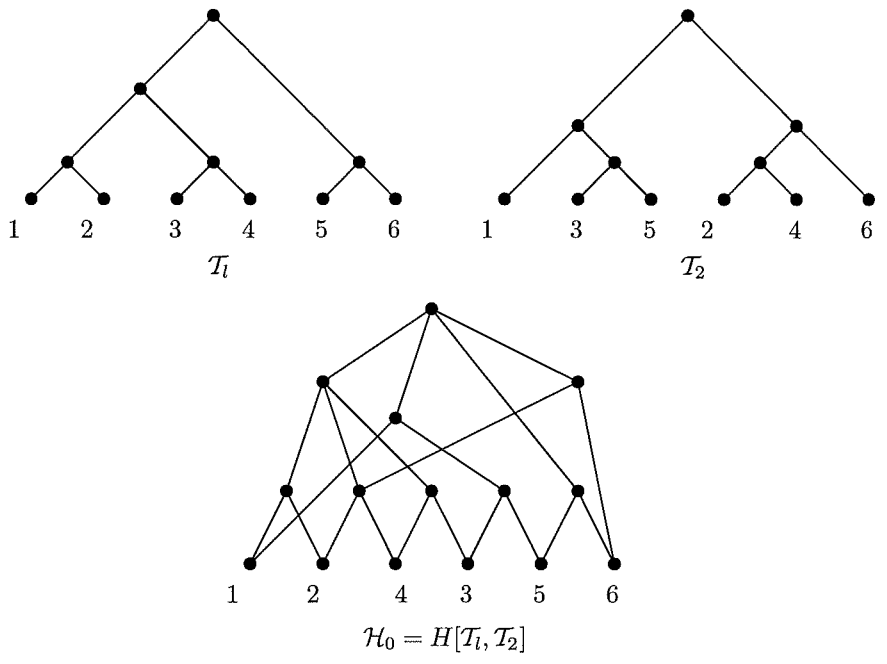


Figure 4.16: Two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ and the cluster union hybrid $\mathcal{H}_0$.

$$
\begin{aligned}
q_0 &= (12, 24, 34, 35, 56, 135, \underline{\mathbf{246}}, \mathbf{1234}) \\
q_1 &= (12, 24, 34, 35, \underline{\mathbf{56}}, \underline{\mathbf{135}}, 12346, 1234) \\
q_2 &= (12, \underline{\mathbf{24}}, \mathbf{34}, 35, 1356, 1356, 12346, 1234) \\
q_3 &= (\mathbf{12}, \underline{\mathbf{234}}, 34, 35, 1356, 1356, 12346, 1234) \\
q_4 &= (12, 1234, 34, 35, 1356, 1356, 12346, 1234)
\end{aligned}
$$

$$
\begin{aligned}
q'_0 &= (12, 24, 34, 35, 56, 135, \mathbf{246}, \underline{\mathbf{1234}}) \\
q'_1 &= (12, 24, 34, 35, \mathbf{56}, \underline{\mathbf{135}}, 246, 12346) \\
q'_2 &= (12, 24, 34, \underline{\mathbf{35}}, \underline{\mathbf{56}}, 1356, 246, 12346) \\
q'_3 &= (12, 24, \underline{\mathbf{34}}, 356, 356, 1356, \underline{\mathbf{246}}, 12346) \\
q'_4 &= (12, 1234, 2346, 356, 356, 1356, 2346, 12346)
\end{aligned}
$$

Figure 4.17: The changing of clusters described by a sequence of vectors: the sequence q_0–q_4 corresponds to the sequence of hybrids $\mathcal{H}_1$–$\mathcal{H}_4$; q_0–q_4 corresponds to the hybrids $\mathcal{H}'_1$–$\mathcal{H}'_4$. At each step, the operation is applied to the clusters in bold font, the clusters that are replaced are underlined.

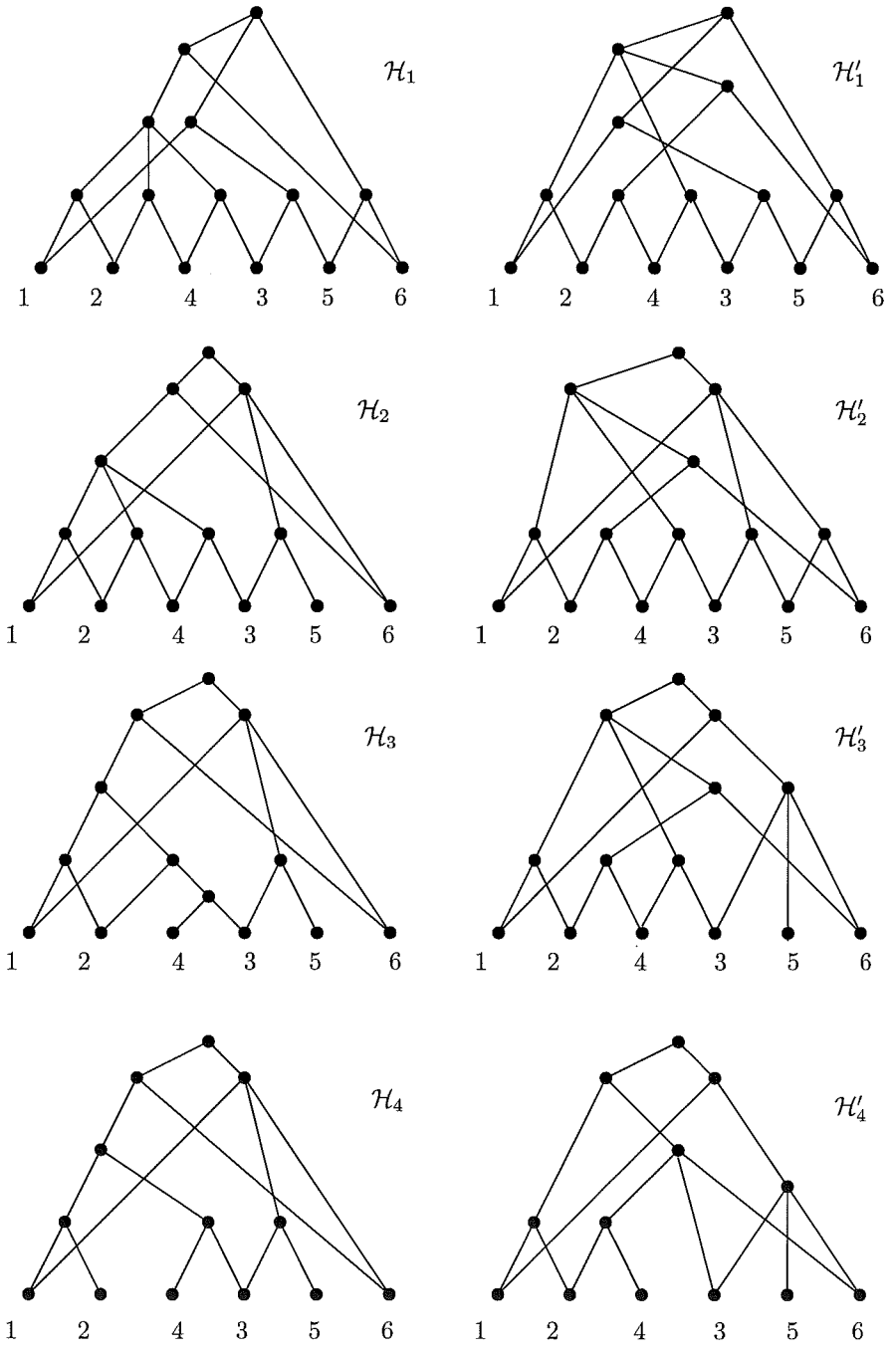


Figure 4.18: Two possible sequences of hybrids $\mathcal{H}_1$–$\mathcal{H}_4$, and respectively $\mathcal{H}'_1$–$\mathcal{H}'_4$, obtained by succesively applying the operation, starting with $\mathcal{H}_0$. The hybrid $\mathcal{H}_4$ is a minimal one; the hybrid $\mathcal{H}'_4$ is not.

Proposition 4.7.1. *Let $\mathcal{T}$ be a rooted phylogenetic X–tree and $\mathcal{H}$ a regular hybrid phylogeny on X. If $c(\mathcal{H})$ properly displays $c(\mathcal{T})$ then $\mathcal{H}$ displays $\mathcal{T}$.*

Proof. For each $C \in c(\mathcal{T})$ consider the corresponding $Q_C \in c(\mathcal{H})$, and for every $(P, C) \in L_{c(T)}$ consider a chain $K(Q_P, Q_C) : Q_C \subseteq A_1 \subseteq \cdots \subseteq A_i \subseteq Q_P$ in $c(\mathcal{H})$. Then the cover digraph of the cluster set $\{Q : Q \in K(Q_P, Q_C), (P, C) \in L_{c(T)}\}$ is a rooted subdigraph of $\mathcal{H}$ that is a refinement of $\mathcal{T}$. It follows that $\mathcal{H}$ displays $\mathcal{T}$. □

Now observe that if $\mathcal{T}$ is a rooted phylogenetic X–tree and $\mathcal{H}$ is a hybrid on X such that $c(\mathcal{T}) \subseteq c(\mathcal{H})$, then $c(\mathcal{H})$ properly displays $c(\mathcal{T})$, and therefore $\mathcal{H}$ displays $\mathcal{T}$. Indeed, if $c(\mathcal{T}) \subseteq c(\mathcal{H})$ then $Q_C = C$ for each $C \in c(\mathcal{T})$. It follows that Q_C is well-defined, and property (P2) is satisfied. To prove that (P3) holds, let $(P_i, C_i) \in L_{c(T)}$, $i = 1, 2$ with $C_1 \neq C_2$. Since P_i, C_i are clusters of a tree, the following cases are possible:

(1) $C_2 \subset P_2 \subseteq C_1 \subset P_1$,
(2) $C_1 \subset P_1 \subseteq C_2 \subset P_2$,
(3) $P_1 = P_2$, and
(4) $P_1 \cap P_2 = \emptyset$.

It is easily seen that if there exists $Q \in c(\mathcal{H})$ such that $C_1 \cup C_2 \subseteq Q \subseteq P_1 \cap P_2$ then $Q \in \{C_1, P_1\} \cap \{C_2, P_2\}$.

As a consequence, if $\mathcal{T}_1$ and $\mathcal{T}_2$ are two rooted phylogenetic X–trees and $\mathcal{H}$ is a regular hybrid phylogeny on X, then $c(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]) = c(\mathcal{T}_1) \cup c(\mathcal{T}_2)$ properly displays $c(\mathcal{T}_1)$ and $c(\mathcal{T}_2)$.

Now we can define our operation. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted phylogenetic X–trees and denote by $\mathcal{C}_i = c(\mathcal{T}_i)$. Let $\mathcal{C}$ be a cluster system on X that properly displays $\mathcal{C}_1$ and $\mathcal{C}_2$ and satisfies the following property:

$$(P4) \quad \forall R \in \mathcal{C} \text{ there exists } C \in c(\mathcal{T}_1) \cup c(\mathcal{T}_2) \text{ such that } R = Q_C.$$

For $R \in \mathcal{C}$ denote by $\hat{R}$ the set of parents of R in $\mathcal{C}$.

Suppose there exist $C_1 \in \mathcal{C}_1$, $C_2 \in \mathcal{C}_2$ such that

- $Q_{C_2'} \neq Q_{C_1}$ $\forall C_2' \in \mathcal{C}_2$ and $Q_{C_1'} \neq Q_{C_2}$ $\forall C_1' \in \mathcal{C}_1$
  (Q_{C_i} is used by only one of the trees.)

- $\hat{Q}_{C_1} \cap \hat{Q}_{C_2} \neq \emptyset$
  (Q_{C_1} and Q_{C_2} have at least one parent in common.)

Modify $\mathcal{C}$ as follows. For C_1 and C_2 as above:

(i) If $\forall D \in \hat{Q}_{C_1} \cap \hat{Q}_{C_2}$, $Q_{C_1} \cup Q_{C_2} \subsetneq D$ (for any common parent the union is a strict subset) then add $Q_{C_1} \cup Q_{C_2}$. In addition, if $\forall D \in \hat{Q}_{C_i}$ we have $Q_{C_1} \cup Q_{C_2} \subsetneq D$ and if $Q_{C_1} \cup Q_{C_2}$ has been added, then remove at least one of Q_{C_i}, $i = 1, 2$.

(ii) If $Q_{C_1} \cup Q_{C_2} \in \hat{Q}_{C_1} \cap \hat{Q}_{C_2}$, and $Q_{C_1} \cup Q_{C_2} \neq Q_C$, $\forall C \in \mathcal{C}_1$ (respectively $\mathcal{C}_2$), (the union is used by only one of the trees), then remove Q_{C_2} (respectively Q_{C_1}).

By applying operations (i)–(ii), the obtained collection $\mathcal{C}'$ of clusters properly displays the cluster system of each tree and (P4) is satisfied. It follows that the regular hybrid $\mathcal{H}'$ corresponding to $\mathcal{C}'$ displays both $\mathcal{T}_1$ and $\mathcal{T}_2$.

If we denote by $\mathcal{H}$ the regular hybrid with the cluster set $\mathcal{C}$, we can observe that, by applying the operations (i)–(ii), $h(\mathcal{H})$ could increase. Indeed, let Q be a cluster of $\mathcal{C}$ (corresponding to a vertex of $\mathcal{H}$), and P and C a parent (respectively a child) of Q in $\mathcal{C}$. By removing Q, in the new hybrid $\mathcal{H}'$, the arcs corresponding to (P, Q) and (Q, C) are removed and if there is no path from P to C in $\mathcal{H}$ that does not contain Q then a new arc (P, C) is added. If we denote by m the number of parents of Q, by l the number of children, and by p the number of pairs (P, C) such that there exists a path from P to C not containing Q, then $|A(\mathcal{H}')| = |A(\mathcal{H})| + ml - (m + l) - p$. Also $|V(\mathcal{H}')| = |V(\mathcal{H})| - 1$. Therefore, if $ml - (m + l) - p - 1 \geq 0$ then $h(\mathcal{H}') \geq h(\mathcal{H})$.

We can use this operation to sometimes obtain $h_r(\mathcal{T}_1, \mathcal{T}_2)$ in the following way.

Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary phylogenetic X–trees. Denote by $\mathcal{C}_i = c(\mathcal{T}_i)$ and let $\mathcal{H}$ be the cluster union hybrid $\mathcal{H} = \mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$. Then $c(\mathcal{H})$ properly displays $\mathcal{C}_i$ and also property (P4) is satisfied.

Start with the cluster system $c(\mathcal{T}_1) \cup c(\mathcal{T}_2)$. Apply the operation until either the hybridization number of the corresponding hybrid increases or the operation cannot

be applied (no pair of clusters with the required properties is found). The obtained regular hybrid $\mathcal{H}'$ displays the two trees and $h(\mathcal{H}') \leq h(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2])$.

The changing of clusters can be described by a sequence of vectors in $\mathbf{R}^n$, where $n = |c(\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]) - X_{triv}|$. See Figure 4.17.

Unfortunately, $h_r(\mathcal{T}_1, \mathcal{T}_2)$ cannot always be obtained. The sequence of regular hybrids depends on what cluster is kept when applying operation (i). Consider for example the two trees in Figure 4.16. Two possible sequences of hybrids $\mathcal{H}_1$–$\mathcal{H}_4$, and respectively $\mathcal{H}'_1$–$\mathcal{H}'_4$, obtained by succesively applying the operation, starting with $\mathcal{H}_0$, are drawn in Figure 4.18. The hybrid $\mathcal{H}_4$ is a minimal one; the hybrid $\mathcal{H}'_4$ is not.

Furthermore, it can happen that none of the regular hybrid phylogenies $\mathcal{H}$ with $h(\mathcal{H}) = h_r(\mathcal{T}_1, \mathcal{T}_2)$ properly displays $\mathcal{T}_1$ and $\mathcal{T}_2$, so $\mathcal{H}$ could not be obtained by the process we described. For example, the hybrid $\mathcal{H}$ in Figure 4.19 is a minimal hybrid that displays the two trees $\mathcal{T}_1$ and $\mathcal{T}_2$, but $\mathcal{T}_2$ is not properly displayed. Also, in this example, property (P4) is not satisfied (the cluster $\{2, 6\}$ of the hybrid does not represent any cluster of the two trees).

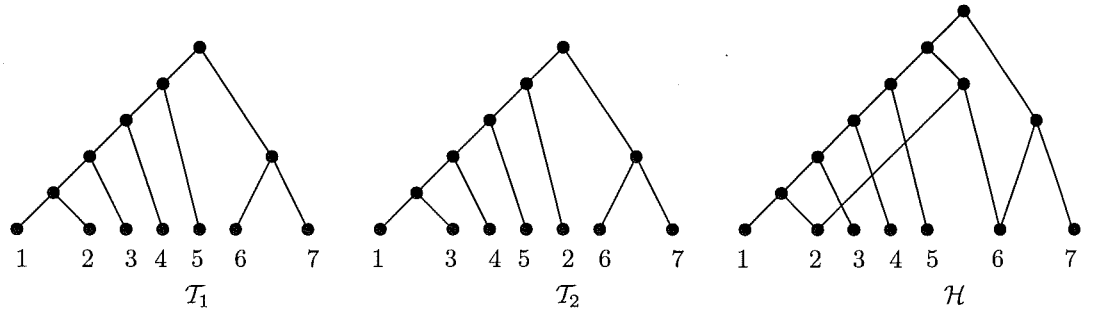


Figure 4.19: Two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ and a minimal hybrid $\mathcal{H}$ that displays the two trees but cannot be obtained by applying the operation.

4.8 Concluding comments and questions for future work

We have proved that

$$d_{rSPR}(\mathcal{T},\mathcal{T}') \leq h(\mathcal{T},\mathcal{T}') \leq n-2,$$

and also that

$$h_r^+(\mathcal{T},\mathcal{T}') = h(\mathcal{T},\mathcal{T}') \leq h_r(\mathcal{T},\mathcal{T}') \leq h(\mathcal{H}[\mathcal{T},\mathcal{T}']) \leq 3n-6.$$

- It would be desirable to find better upper bounds for $h_r(\mathcal{T},\mathcal{T}')$. We conjecture that $h_r(\mathcal{T},\mathcal{T}') \leq n-2$.

- Taking into account the relationship between $h(\mathcal{T},\mathcal{T}')$ and $d_{rSPR}(\mathcal{T},\mathcal{T}')$ (see also Chapter 5), we conjecture that computing $h(\mathcal{T},\mathcal{T}')$ is an NP–hard problem. What about the complexity of computing $h_r(\mathcal{T},\mathcal{T}')$?

- Consider the 'vertex–sharing' type of display (let us call it vs–display) introduced in Section 3.8 and define

  $$h_{vs}(\mathcal{T},\mathcal{T}') = \min\{h(\mathcal{H}) : \mathcal{H} \text{ a hybrid that vs–display } \mathcal{T} \text{ and } \mathcal{T}'\}.$$

  Note that for the trees in Figure 3.23, $h_{vs}(\mathcal{T},\mathcal{T}') = 1 < 2 = h(\mathcal{T},\mathcal{T}')$. We conjecture that $h_{vs}(\mathcal{T},\mathcal{T}') \leq h(\mathcal{T},\mathcal{T}')$, for any pair of rooted binary trees $\mathcal{T}$ and $\mathcal{T}'$. What is the relationship between $h_{vs}(\mathcal{T},\mathcal{T}')$ and $d_{rSPR}(\mathcal{T},\mathcal{T}')$?

- In Section 4.6 we have constructed pairs of rooted binary trees $\mathcal{T}$ and $\mathcal{T}'$ for which the difference between $h(\mathcal{T},\mathcal{T}')$ and $d_{rSPR}(\mathcal{T},\mathcal{T}')$ is large. A further question is the following: For a given n, determine

  $$\mu(n) = \inf\{h(\mathcal{T},\mathcal{T}') - d_{rSPR}(\mathcal{T},\mathcal{T}') : \mathcal{T},\mathcal{T}' \in RB(n)\}$$

  or at least find better bounds for $\mu(n)$.

- It follows from [51] and Section 4.6 that $n/2 \leq \mathrm{diam}_{rSPR}RB(n) \leq n-2$. Can these bounds be improved?

- Develop computational tools based on the model of hybrid phylogenies.

Chapter 5

How to construct a minimal hybrid—an example from biology

The aim of this chapter is to show how the graph–theoretic framework developed in Chapter 4, particularly good agreement forests, can be used for analysing and representing hybrid evolution.

5.1 How to construct a minimal hybrid

In Chapter 4, we showed that for any pair of rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$,

$$d_{rSPR}(\mathcal{T}, \mathcal{T}') \leq m_g(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}, \mathcal{T}').$$

For practical reasons, it is useful to have an algorithm that constructs an 'appropriate – size' hybrid that displays $\mathcal{T}$ and $\mathcal{T}'$. In this section, we describe two such algorithms: one that constructs the hybrid starting from a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, and the other starting from a sequence of rSPR operations associated with a good agreement forest of the two trees.

First, let us observe that there is a correspondence between sequences of rSPR operations that transform $\mathcal{T}$ into $\mathcal{T}'$ and agreement forests for the two trees. This

correspondence is not one–to–one; the same agreement forest can be associated to different sequences of rSPR operations. For example the two sequences of rSPR operation from Figure 5.1 have the same agreement forest.

Given a sequence of p rSPR operations

$$\mathcal{T} = \mathcal{T}_0 \overset{\theta_1}{\to} \mathcal{T}_1 \overset{\theta_2}{\to} \mathcal{T}_2 \overset{\theta_3}{\to} \cdots \overset{\theta_{i-1}}{\to} \mathcal{T}_{i-1} \overset{\theta_i}{\to} \mathcal{T}_i \overset{\theta_{i+1}}{\to} \mathcal{T}_{i+1} \overset{\theta_{i+2}}{\to} \cdots \overset{\theta_p}{\to} \mathcal{T}_p = \mathcal{T}',$$

we can obtain an agreement forest in the following way (see also [11]).

Step (1). Let $\{A \cup \rho, B\}$ be the partition of $X \cup \{\rho\}$ induced by θ_1. Then $\{\mathcal{T}_\rho^1, \mathcal{T}_1^1\}$, where $\mathcal{T}_\rho^1 = \mathcal{T}|(A \cup \{\rho\})$, $\mathcal{T}_1^1 = \mathcal{T}|B$, is an agreement forest for $\mathcal{T}$ and $\mathcal{T}_1$. Note that this is always a good agreement forest.

Step (i). Suppose we have constructed an agreement forest

$$\{\mathcal{T}_\rho^{i-1}, \mathcal{T}_1^{i-1}, \mathcal{T}_2^{i-1}, \ldots, \mathcal{T}_j^{i-1}\}$$

for $\mathcal{T}$ and $\mathcal{T}_{i-1}$. Also, as in step 1, we construct an agreement forest $\{\mathcal{T}_\rho^i, \mathcal{T}_1^i\}$ for $\mathcal{T}_{i-1}$ and $\mathcal{T}_i$. The partition $\mathcal{L}(\mathcal{T}_\rho^i), \mathcal{L}(\mathcal{T}_1^i)$ identifies a unique edge in $\mathcal{T}_{i-1}$, hence there exists at most one $k \in \{\rho, 1, 2, \ldots, j\}$ such that $\mathcal{L}(\mathcal{T}_k^{i-1}) \cap \mathcal{L}(\mathcal{T}_\rho^i) \neq \emptyset$ and $\mathcal{L}(\mathcal{T}_k^{i-1}) \cap \mathcal{L}(\mathcal{T}_1^i) \neq \emptyset$. If there is no such k then $\{\mathcal{T}_\rho^{i-1}, \mathcal{T}_1^{i-1}, \mathcal{T}_2^{i-1}, \ldots, \mathcal{T}_j^{i-1}\}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}_i$. If such k exists then denote by $\mathcal{L}_{k,\rho} = \mathcal{L}(\mathcal{T}_k^{i-1}) \cap \mathcal{L}(\mathcal{T}_\rho^i)$ and by $\mathcal{L}_{k,1} = \mathcal{L}(\mathcal{T}_k^{i-1}) \cap \mathcal{L}(\mathcal{T}_1^i)$. It follows that

$$\{\mathcal{T}_s^{i-1} : s \in \{\rho, 1, \ldots, j\} - \{k\}\} \cup \{\mathcal{T}_{i-1}|\mathcal{L}_{k,\rho}, \mathcal{T}_{i-1}|\mathcal{L}_{k,1}\}$$

is an agreement forest for $\mathcal{T}$ and $\mathcal{T}_i$.

Finally, the agreement forest $\mathcal{F}$ obtained in step p is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Note that $\mathcal{F}$ is also an agreement forest for any pair $(\mathcal{T}_i, \mathcal{T}_j)$ of trees in the sequence.

We show now how a minimal hybrid can be constructed starting from a good agreement forest. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X–trees and let $\mathcal{F}_k = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. We can use the proof of Theorem 4.3.2, to construct a minimal hybrid that displays both trees. As $\mathcal{F}_k$ is good, $\mathcal{G}_{\mathcal{F}_k}$ has no directed cycles; whence there exists a vertex of

$\mathcal{G}_{\mathcal{F}_k}$ that has out–degree zero. Without loss of generality, we may assume that this vertex is $\mathcal{T}_k$. Then $\mathcal{F}_{k-1} = \mathcal{F}_k - \{\mathcal{T}_k\}$ is a good agreement forest for $\mathcal{T}|X_{k-1}$ and $\mathcal{T}'|X_{k-1}$, where X_{k-1} denotes $X - \mathcal{L}(\mathcal{T}_k)$. It follows that $\mathcal{G}_{\mathcal{F}_{k-1}}$ contains a vertex of out–degree zero, say $\mathcal{T}_{k-1}$, and so on. For $1 \leq i \leq k$, denote by $\mathcal{F}_{i-1} = \mathcal{F}_i - \{\mathcal{T}_i\}$ and by $X_{i-1} = X_i - \mathcal{L}(\mathcal{T}_i)$. It follows that $\mathcal{F}_0 = \{\mathcal{T}_\rho\}$ and $X_0 = \mathcal{L}(\mathcal{T}_\rho)$.

- Given $\mathcal{T}$, $\mathcal{T}'$ and a maximum good agreement forest $\mathcal{F}_k$ as above, construct $\mathcal{G}_k = \mathcal{G}_{\mathcal{F}_k}$.
- Choose a vertex $\mathcal{T}_k$ of out–degree zero and consider $\mathcal{G}_{k-1} = \mathcal{G}_k - \mathcal{T}_k$. Repeat the operation until $\mathcal{G}_0 = \mathcal{T}_\rho$ is obtained.

Now we will construct a minimal hybrid starting from $\mathcal{T}_0 = \mathcal{T}_\rho$ and succesively adjoining the trees $\mathcal{T}_1, \ldots, \mathcal{T}_k$.

- Start with $\mathcal{H}_0 = \mathcal{T}_\rho$.
- Construct the hybrid $\mathcal{H}_1$ from $\mathcal{H}_0$ by adjoining the root $\mathcal{T}_1$ using the edges e_1 and e_1' such that $\mathcal{H}_1$ displays $\mathcal{T}|(X_0 \cup X_1)$ and respectively $\mathcal{T}'|(X_0 \cup X_1)$.
- Repeat the construction for $2 \leq i \leq k$. Adjoin succesively the trees $\mathcal{T}_2, \ldots, \mathcal{T}_k$ and correspondingly obtain the hybrids $\mathcal{H}_2, \ldots, \mathcal{H}_k$.

It is easily seen that in each step the hybridization number increases by one, for otherwise $\mathcal{F}_k$ is not a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Finally, the hybrid $\mathcal{H} = \mathcal{H}_k$ has $h(\mathcal{H}) = k$ and the hybridization vertices are the roots of adjoined trees.

An example of this construction is given in Figure 5.1(b); the hybrid is obtained by applying the previous construction to a maximum good agreement forest for the trees $\mathcal{T}$ and $\mathcal{T}'$ drawn in the same figure.

Note that the hybrid obtained by applying the above construction is not unique. The construction depends on the order in which we attach the subtrees. (There can exist more vertices with out–degree zero in $\mathcal{G}_i$, so the order depends on the vertex we choose.)

In Proposition 4.4.1 we have shown how one can construct a regular hybrid that displays two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ for which $d_{rSPR}(\mathcal{T}, \mathcal{T}') = 1$. If appropriate sequences of rSPR operations are considered, this construction can be adapted for the case $d_{rSPR} \geq 2$.

The fact that not all sequences of rSPR operations are appropriate for reconstructing evolutionary histories is not surprising. As pointed out by Maddison [36]: "What is needed is a method that counts the minimal number of branch moves needed to convert one tree into another, where branch moves are restricted so as not to violate a linear time order (one can imagine a series of branch moves that cannot possibly happen together, e.g., one move from branch A to branch B and then another move from a descendant of B to an ancestor of A)."

For other approaches to the problem of reconstructing phylogenetic networks from sequences of rSPR operations, see for example [40, 52].

We now show how a minimal hybrid can be constructed starting from a sequence of rSPR operations.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic trees such that $d_{rSPR}(\mathcal{T}, \mathcal{T}') = k$. Then there exists a sequence θ_i, $1 \leq i \leq k$, of rSPR operations:

$$\mathcal{T} = \mathcal{T}_0 \overset{\theta_1}{\rightarrow} \mathcal{T}_1 \overset{\theta_2}{\rightarrow} \mathcal{T}_2 \overset{\theta_3}{\rightarrow} \cdots \overset{\theta_{i-1}}{\rightarrow} \mathcal{T}_{i-1} \overset{\theta_i}{\rightarrow} \mathcal{T}_i \overset{\theta_{i+1}}{\rightarrow} \mathcal{T}_{i+1} \overset{\theta_{i+2}}{\rightarrow} \cdots \overset{\theta_k}{\rightarrow} \mathcal{T}_k = \mathcal{T}',$$

where for each $1 \leq i \leq k$, θ_i is a rSPR operation that prunes the subtree S_{i-1} with the root a_{i-1} of $\mathcal{T}_{i-1}$ and regrafts it up to the vertex c_{i-1}, by subdividing the arc incident to c_{i-1}. Construct a rooted digraph as follows:

- Start with $\mathcal{D}_0 = \mathcal{T}_0 = \mathcal{T}$.
- Step(1). Let $\mathcal{D}_1$ be the digraph obtained from $\mathcal{D}_0$ in the following way: Subdivide the arc incident to c_0 by a new vertex v_0 and add a new arc from v_0 to a_0. Denote by p_0 the arc incident to a_0 in $\mathcal{T}_0$. Then it is easily seen that $\mathcal{D}_1 - p_0$ is a refinement of $\mathcal{T}_1$, and $\mathcal{D}_1$ is a hybrid that displays both $\mathcal{T}_0$ and $\mathcal{T}_1$.
- Step(i). Suppose that $\mathcal{D}_{i-1}$ has been constructed. Let $\mathcal{T}_i$ be the tree obtained from $\mathcal{T}_{i-1}$ by applying the operation θ_i. Construct a digraph $\mathcal{D}_i$ by modifying

$\mathcal{D}_{i-1}$ as follows. Subdivide the arc incident to c_{i-1} (corresponding to the unique arc in the tree on which the rSPR operation was being performed) by a new vertex v_{i-1}, and add a new arc (v_{i-1}, a_{i-1}). Let p_{i-1} be the set of arcs that subdivides the arc incident with a_{i-1} in $\mathcal{T}_{i-1}$. Then $\mathcal{D}_i - \{p_0, p_1, \ldots p_{i-1}\}$ is a refinement of $\mathcal{T}_i$.

After k steps, the obtained digraph $\mathcal{D} = \mathcal{D}_k$ contains refinements of $\mathcal{T}_0$, $\mathcal{T}_1$, $\ldots \mathcal{T}_k$ as subdigraphs. If $\mathcal{D}$ is acyclic, then $(\mathcal{D}, \psi)$, where ψ is the leaf labelling of $\mathcal{T}_0$, is a hybrid phylogeny that displays all the trees in the sequence (in particular, displays $\mathcal{T}_0$ and $\mathcal{T}_k$), and $h(\mathcal{D}) = k$.

An example of this construction is the hybrid drawn in Figure 5.1(a). The added arcs are represented by arrows, and the arcs on the paths p_i are shown by dashed lines.

The above construction can induce directed cycles, so if this is the case, the obtained digraph is not a hybrid. An example is given in Figure 5.3 a). Note that the agreement forest corresponding to the sequence of rSPR operations used in the construction of the digraph drawn in Figure 5.3 is not a good agreement forest. Also, the hybrid in Figure 5.3 b) is obtained by applying the construction to a sequence of four rSPR operations that has attached a good agreement forest. As we will show later, this fact is not a coincidence.[1]

Denote by $\overline{d}_{rSPR}(\mathcal{T}, \mathcal{T}')$ the minimum number of rSPR operations required to transform $\mathcal{T}$ into $\mathcal{T}'$ such that the above construction does not induce cycles.

It follows that $m_g(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}, \mathcal{T}') \leq \overline{d}_{rSPR}(\mathcal{T}, \mathcal{T}')$. The following proposition shows the relation between good agreement forests and sequences of rSPR operations appropriate for constructing hybrid phylogenies.

Proposition 5.1.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X–trees. Then*

$$\overline{d}_{rSPR}(\mathcal{T}, \mathcal{T}') = m_g(\mathcal{T}, \mathcal{T}').$$

[1] Actually, the idea of introducing the notion of good agreement forest was inspired by this 'bad' digraph.

Proof. We have to prove that $\overline{d}_{rSPR}(\mathcal{T}, \mathcal{T}') \leq m_g(\mathcal{T}, \mathcal{T}')$. The proof is by induction on $m_g(\mathcal{T}, \mathcal{T}')$.

Clearly, the inequality holds for $m_g(\mathcal{T}, \mathcal{T}') = 0$. Now assume the result holds for all pairs of rooted binary phylogenetic X–trees for which the minimum number of components over all good agreement forests is at most k, and let $m_g(\mathcal{T}, \mathcal{T}')) = k$. Then there exists $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \ldots, \mathcal{T}_k\}$ a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Using the same ideas as in the proof of Theorem 4.3.2, it follows that one of the trees in $\mathcal{F}$, say $\mathcal{T}_k$, is a rooted subtree of both $\mathcal{T}$ and $\mathcal{T}'$. Then $\mathcal{F}_k = \mathcal{F} - \{\mathcal{T}_k\}$ is a good agreement forest for $\mathcal{T}|(X - \mathcal{L}(X_k))$ and $\mathcal{T}'|(X - \mathcal{L}(X_k))$, with $|\mathcal{F}_k| < |\mathcal{F}|$.

By induction assumption, it follows that there is a sequence of at most $k-1$ rSPR operations from $\mathcal{T}|(X-\mathcal{L}(X_k))$ to $\mathcal{T}'|(X-\mathcal{L}(X_k))$ such that the corresponding digraph $\mathcal{H}_k$ is a hybrid on $X - \mathcal{L}(X_k)$. By applying the same sequence of operations to $\mathcal{T}$, we obtain an X–tree $\mathcal{T}''$ containing a refinement of $\mathcal{T}'|(X - \mathcal{L}(X_k))$ and $\mathcal{T}_k$ as subtrees. Also, by applying to $\mathcal{T}$ the construction corresponding to this sequence, we obtain a hybrid $\mathcal{H}$ on X that displays $\mathcal{T}$ and $\mathcal{T}''$, and contains $\mathcal{H}_k$ and $\mathcal{T}_k$ as subdigraphs. Moreover, there is no directed path in $\mathcal{H}$ from a vertex of $\mathcal{T}_k$ to a vertex of $\mathcal{H}_k$.

Now, let us observe that $\mathcal{T}'$ can be obtained from $\mathcal{T}''$ by a single rSPR operation that prunes $\mathcal{T}_k$ and reattaches it to the correct place. Corresponding to this operation, a new arc is added to $\mathcal{H}$, from a vertex that subdivides an arc in $\mathcal{H}_k$ to the root of $\mathcal{T}_k$. The obtained digraph corresponds to a sequence of at most k rSPR operations and is acyclic. The conclusion now follows. □

Consequently, we have that

$$h(\mathcal{T}, \mathcal{T}') = \overline{d}_{rSPR}(\mathcal{T}, \mathcal{T}').$$

Note that the hybrid obtained starting from a sequence of rSPR operations associated to a maximum good agreement forest $\mathcal{F}$ might be different from the hybrid constructed directly from $\mathcal{F}$. (See Figure 5.1 for an example.) Also, let us observe that the hybrid, $\mathcal{H}$ say, obtained by either of the constructions is not necessarily regular. However, if a regular hybrid is needed, we can apply the construction indicated in Proposition 3.5.1 (add extra–leaves) to $\mathcal{H}$, and obtain a regular hybrid with the same hybridization number.

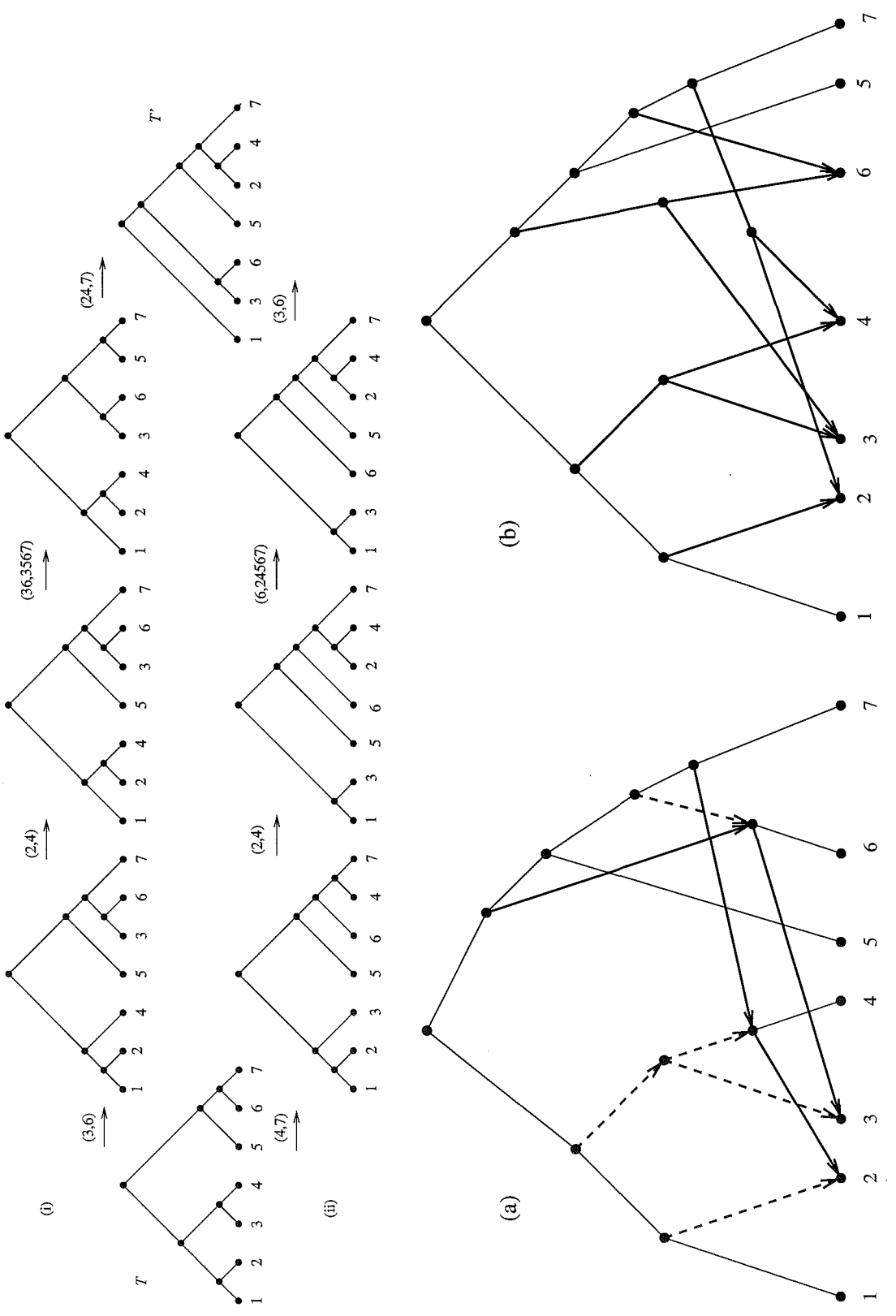


Figure 5.1: Two sequences of rSPR operations (i) and (ii) that transform $\mathcal{T}$ into $\mathcal{T}'$ and two hybrids that display both trees.

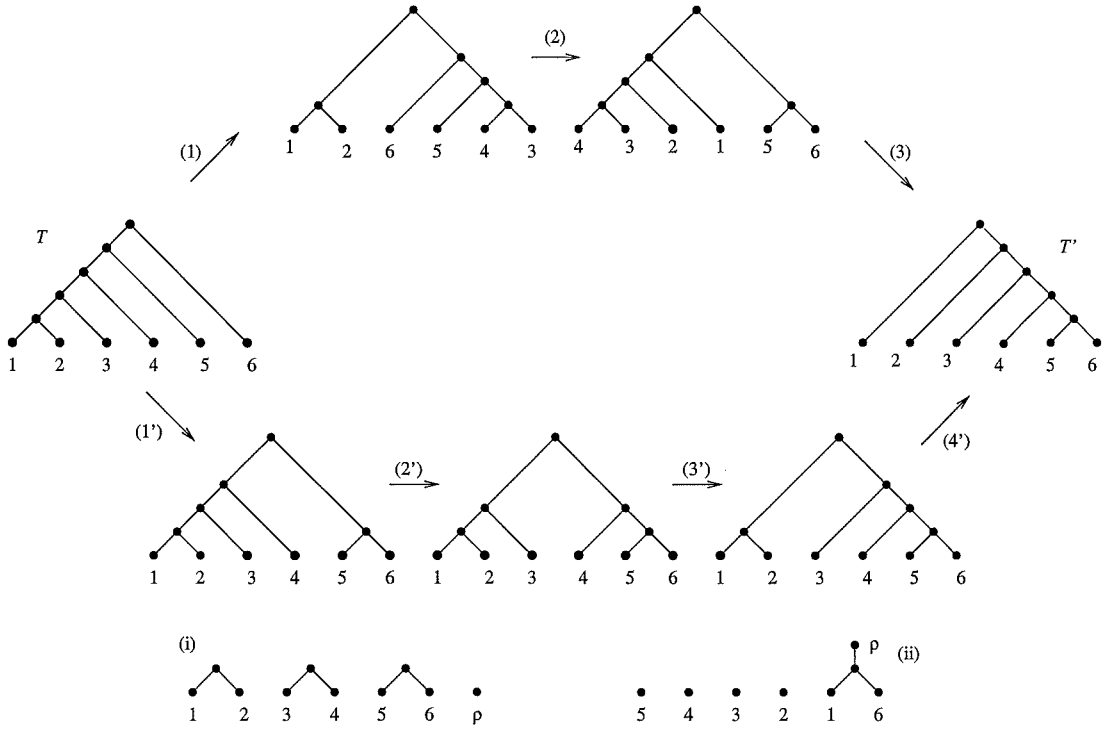


Figure 5.2: Two sequences of rSPR operations and the associated agreement forests: the sequence (1)–(3) corresponds to $d_{rSPR}(\mathcal{T}, \mathcal{T}')$; (1')–(4') corresponds to $h(\mathcal{T}, \mathcal{T}')$.

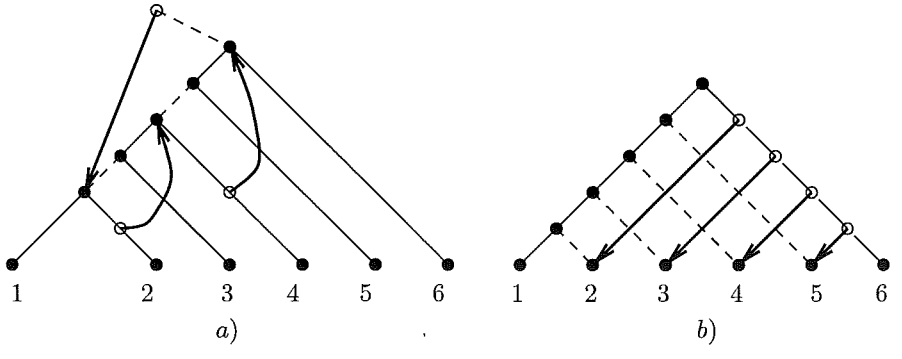


Figure 5.3: Two digraphs corresponding to the sequences of rSPR operations represented in Figure 5.2: a) The digraph $\mathcal{D}$ corresponding to the sequence (1)–(3) is not a hybrid as it contains cycles. b) A minimal hybrid corresponding to the sequence (1')–(4') of rSPR operations. Note that it displays the initial and the final trees as well as all the trees in the sequence.

5.2 'Reducing' the problem

Since computing the rSPR distance between two rooted binary phylogenetic X–trees is NP–hard [11], and taking into account the results in Section 5.1, it seems very likely that computing $h(\mathcal{T}, \mathcal{T}')$ (and finding a maximum good agreement forest) is also NP–hard.

Confronted with the problem of computing $h(\mathcal{T}, \mathcal{T}')$, one can try to reduce the problem to 'small' trees, by considering common clusters for $\mathcal{T}$ and $\mathcal{T}'$. We next consider this problem. First, we consider the restrictions of the two X–trees to a common cluster A and to $X - A$.

Proposition 5.2.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees, and suppose that there exists a cluster $A \in c(\mathcal{T}) \cap c(\mathcal{T}')$, $A \neq X$. Let $\overline{A} = X - A$. Then*

$$h(\mathcal{T}, \mathcal{T}') - 1 \leq h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}|\overline{A}, \mathcal{T}'|\overline{A}) \leq h(\mathcal{T}, \mathcal{T}').$$

Proof. For the first inequality, let $\mathcal{F}_A = \{\mathcal{T}_\rho^A, \mathcal{T}_1^A, \ldots, \mathcal{T}_k^A\}$ be a maximum good agreement forest for $\mathcal{T}|A, \mathcal{T}'|A$ and $\mathcal{F}_{\overline{A}} = \{\mathcal{T}_\rho^{\overline{A}}, \mathcal{T}_1^{\overline{A}}, \ldots, \mathcal{T}_s^{\overline{A}}\}$ be a maximum good agreement forest for $\mathcal{T}|\overline{A}, \mathcal{T}'|\overline{A}$. We will prove that $\mathcal{F} = \{\mathcal{T}_\rho^A|A, \mathcal{T}_1^A, \ldots, \mathcal{T}_k^A\} \cup \mathcal{F}_{\overline{A}}$ is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.

Clearly, $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Denote by r_i the root of $\mathcal{T}(\mathcal{L}(\mathcal{T}_i^A))$ and by $\overline{r_j}$ the root of $\mathcal{T}(\mathcal{L}(\mathcal{T}_j^{\overline{A}}))$. Assume now that $\mathcal{F}$ is not a good agreement forest. Therefore there exists a directed cycle in $\mathcal{G}_\mathcal{F}$. Since $\mathcal{F}_A$ and $\mathcal{F}_{\overline{A}}$ do not contain cycles, it follows that this cycle contains at least one vertex r_i and one vertex $\overline{r_j}$. Then there exists a directed path in $\mathcal{G}_\mathcal{F}$ from r_i to $\overline{r_j}$ and a directed path from $\overline{r_j}$ to r_i. Therefore r_i is an ancestor of $\overline{r_j}$ in $\mathcal{T}$ (or $\mathcal{T}'$) and $\overline{r_j}$ is an ancestor of r_i in $\mathcal{T}'$ (or $\mathcal{T}$). It results that A is not a common cluster of both trees, a contradiction.

As in a good agreement forest ρ is not an isolated vertex (Lemma 4.3.1), it follows that $|\mathcal{F}_A| + |\mathcal{F}_{\overline{A}}| = |\mathcal{F}|$.

Therefore,

$$\begin{aligned} h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}|\overline{A}, \mathcal{T}'|\overline{A}) &= |\mathcal{F}_A| - 1 + |\mathcal{F}_{\overline{A}}| - 1 \\ &= |\mathcal{F}| - 1 - 1 \\ &\geq h(\mathcal{T}, \mathcal{T}') - 1. \end{aligned}$$

For the second inequality, consider a maximum good agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$. There are two possible cases to consider:

(i) there exists $\mathcal{T}_i \in \mathcal{F}$ such that $\mathcal{L}(\mathcal{T}_i) \cap A \neq \emptyset$ and $\mathcal{L}(\mathcal{T}_i) \cap (\overline{A} \cup \{\rho\}) \neq \emptyset$, and

(ii) for all $\mathcal{T}_i \in \mathcal{F}$, either $\mathcal{L}(\mathcal{T}_i) \subseteq A$ or $\mathcal{L}(\mathcal{T}_i) \subseteq (\overline{A} \cup \{\rho\})$.

Case (i). If there is a tree $\mathcal{T}_i$ with the properties in case (i), then $\mathcal{T}(\mathcal{L}_i)$ contains the root of $\mathcal{T}|A$. It follows that $\mathcal{T}_i$ is unique with the properties from (i) for otherwise $\mathcal{F}$ is not an agreement forest. Now let $x \in \mathcal{L}(\mathcal{T}_i) \cap (\overline{A} \cup \rho)$ and let $\mathcal{T}_{i,A}$ be the tree obtained from $\mathcal{T}_i|(A \cup x)$ by relabelling x as ρ. Then $\mathcal{F}_A = \{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq A\} \cup \{T_{i,A}\}$ is a good agreement forest for $T|A$ and $T'|A$. Also $\mathcal{F}_{\overline{A}} = \{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq \overline{A} \cup \{\rho\}\} \cup \{\mathcal{T}_i|(\overline{A} \cup \{\rho\})\}$ is a good agreement forest for $\mathcal{T}_{\overline{A}}$ and $\mathcal{T}'_{\overline{A}}$. Hence, $|\mathcal{F}| = |\mathcal{F}_A| + |\mathcal{F}_{\overline{A}}| - 1$, and

$$\begin{aligned} h(\mathcal{T}, \mathcal{T}') &= |\mathcal{F}| - 1 \\ &= |\mathcal{F}_A| + |\mathcal{F}_{\overline{A}} - 1 - 1 \\ &\geq h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}|\overline{A}, \mathcal{T}'|\overline{A}). \end{aligned}$$

Case (ii). Since $\mathcal{G}_{\mathcal{F}}$ does not contain directed cycles, it follows that the subdigraph of $\mathcal{G}_{\mathcal{F}}$ corresponding to the set $\{\mathcal{T}_i \in \mathcal{F} : \mathcal{L}(\mathcal{T}_i) \subseteq A\}$ does not contain directed cycles. Thus, this subdigraph has a vertex of in-degree zero, say $\mathcal{T}_0$. Let $\mathcal{T}_{0,\rho}$ be the tree obtained from $\mathcal{T}_0$ by adding ρ at the end of a pendant edge adjoined to the root of $\mathcal{T}_0$. Then $\mathcal{F}_A = \{\mathcal{T}_i \in \mathcal{F} : \mathcal{L}(\mathcal{T}_i) \subseteq A\} - \{\mathcal{T}_0\} \cup \{\mathcal{T}_{0,\rho}\}$ is a good agreement forest for $\mathcal{T}|A$ and $\mathcal{T}'|A$. Also, $\mathcal{F}_{\overline{A}} = \{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq \overline{A} \cup \{\rho\}\}$ is a good agreement forest for $\mathcal{T}_{\overline{A}}$ and $\mathcal{T}'_{\overline{A}}$. So, in this case, $|\mathcal{F}| = |\mathcal{F}_A| + |\mathcal{F}_{\overline{A}}|$, and

$$\begin{aligned} h(\mathcal{T}, \mathcal{T}') > h(\mathcal{T}, \mathcal{T}') - 1 &= |\mathcal{F}| - 1 - 1 \\ &= |\mathcal{F}_A| + |\mathcal{F}_{\overline{A}}| - 1 - 1 \\ &\geq h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}|\overline{A}, \mathcal{T}'|\overline{A}). \end{aligned}$$

This completes the proof of the proposition. □

The following two corollaries can be used to approximate $h(\mathcal{T}, \mathcal{T}')$.

Corollary 5.2.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees and suppose that there exists a cluster $A \in c(\mathcal{T}) \cap c(\mathcal{T}')$, $A \neq X$, and such that $\mathcal{T}|A = \mathcal{T}'|A$. Let $\overline{A} = X - A$. Then*

$$h(\mathcal{T}, \mathcal{T}') - 1 \leq h(\mathcal{T}|\overline{A}, \mathcal{T}'|\overline{A}) \leq h(\mathcal{T}, \mathcal{T}').$$

In particular, for each $x \in X$ we have:

$$h(\mathcal{T}, \mathcal{T}') - 1 \leq h(\mathcal{T}|(X - \{x\}), \mathcal{T}'|(X - \{x\})) \leq h(\mathcal{T}, \mathcal{T}').$$

Proof. If $\mathcal{T}|A = \mathcal{T}'|A$ then $h(\mathcal{T}|A, \mathcal{T}'|A) = 0$. In particular, $h(\mathcal{T}|\{x\}, \mathcal{T}'|\{x\}) = 0$. The result follows from Proposition 5.2.1. □

Corollary 5.2.3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Then for any $A \subset X$ we have:*

$$h(\mathcal{T}, \mathcal{T}') - |A| \leq h(\mathcal{T}|(X - A), \mathcal{T}'|(X - A)) \leq h(\mathcal{T}, \mathcal{T}').$$

Proof. Let $A = \{x_1, x_2, \ldots, x_k\} \subset X$. For $p \in \{1, 2, \ldots, k\}$, denote by

$$h_p = h(\mathcal{T}|(X - \{x_1, x_2, \ldots, x_p\}), \mathcal{T}'|(X - \{x_1, x_2, \ldots, x_p\})).$$

Observe that, with this notation,

$$h_k = h(\mathcal{T}|(X - A), \mathcal{T}'|(X - A)).$$

Now by succesively applying Corollary 5.2.2 we have:

$$\begin{array}{rcccl} h(\mathcal{T}, \mathcal{T}') - 1 & \leq & h_1 & \leq & h(\mathcal{T}, \mathcal{T}') \\ h_1 - 1 & \leq & h_2 & \leq & h_1 \\ & & \cdots & & \\ h_{k-1} - 1 & \leq & h_k & \leq & h_{k-1}. \end{array}$$

By adding the above inequalities we obtain:

$$h(\mathcal{T}, \mathcal{T}') - k \leq h_k \leq h(\mathcal{T}, \mathcal{T}').$$

□

The next corollary provides an alternative proof of Proposition 4.4.2.

Corollary 5.2.4. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Then*

$$0 \leq h(\mathcal{T}, \mathcal{T}') \leq |X| - 2.$$

Proof. Let $x, y \in X$ and let $A = X - \{x, y\}$. From Corollary 5.2.3, it follows that

$$h(\mathcal{T}, \mathcal{T}') - (|X| - 2) \leq h(\mathcal{T}|\{x, y\}, \mathcal{T}'|\{x, y\}) \leq h(\mathcal{T}, \mathcal{T}').$$

Then observe that $h(\mathcal{T}|\{x, y\}, \mathcal{T}'|\{x, y\}) = 0$. □

The following proposition shows how $h(\mathcal{T}, \mathcal{T}')$ can be obtained by reducing the problem to certain trees with a smaller number of leaves.

Proposition 5.2.5. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees, and suppose that there exists a cluster $A \in c(\mathcal{T}) \cap c(\mathcal{T}')$. Let $\overline{\mathcal{T}}$ and $\overline{\mathcal{T}'}$ be the trees obtained from $\mathcal{T}$ (respectively $\mathcal{T}'$) by replacing the subtree having A as the set of leaves with a new leaf $a \notin X$. Then*

$$h(\mathcal{T}|A, \mathcal{T}'|A) + h(\overline{\mathcal{T}}, \overline{\mathcal{T}'}) = h(\mathcal{T}, \mathcal{T}').$$

Proof. Clearly, the equality holds when $A = X$. Let us discuss now the case $A \neq X$. First, we will prove that $h(\mathcal{T}|A, \mathcal{T}'|A) + h(\overline{\mathcal{T}}, \overline{\mathcal{T}'}) \geq h(\mathcal{T}, \mathcal{T}')$. Denote by $\overline{A} = X - A$. Then $\mathcal{L}(\overline{\mathcal{T}}) = \mathcal{L}(\overline{\mathcal{T}'}) = \overline{A} \cup \{a\}$. Let $\mathcal{F}_A = \{\mathcal{T}_\rho^A, \mathcal{T}_1^A, \ldots, \mathcal{T}_k^A\}$ be a maximum good agreement forest for $\mathcal{T}|A$ and $\mathcal{T}'|A$ and $\mathcal{F}_{\overline{A},a}$ be a maximum good agreement forest for $\overline{\mathcal{T}}$ and $\overline{\mathcal{T}'}$. Since $a \in \mathcal{L}(\overline{\mathcal{T}})$, there exists a tree $\mathcal{T}_{j,a}$ in $\mathcal{F}_{\overline{A},a}$ such that $a \in \mathcal{L}(\mathcal{T}_{j,a})$. Let $\mathcal{T}_a$ be the tree obtained from $\mathcal{T}_{j,a}$ by replacing a with $\mathcal{T}_\rho^A$. Then $\mathcal{F} = \mathcal{F}_A \cup \mathcal{F}_{\overline{A},a} - \{\mathcal{T}_\rho^A, \mathcal{T}_{j,a}\} \cup \{\mathcal{T}_a\}$ is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ and $|\mathcal{F}| = |\mathcal{F}_A| + |\mathcal{F}_{\overline{A},a}| - 1$. It follows that

$$\begin{aligned} h(\mathcal{T}|A, \mathcal{T}'|A) + h(\overline{\mathcal{T}}, \overline{\mathcal{T}'}) &= |\mathcal{F}_A| - 1 + |\mathcal{F}_{\overline{A},a}| - 1 \\ &= |\mathcal{F}| - 1 \\ &\geq h(\mathcal{T}, \mathcal{T}'). \end{aligned}$$

We prove now that $h(\mathcal{T}|A, \mathcal{T}'|A) + h(\overline{\mathcal{T}}, \overline{\mathcal{T}'}) \leq h(\mathcal{T}, \mathcal{T}')$. Let $\mathcal{F}$ be a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. There are two possible cases to consider:

(i) there exists $\mathcal{T}_i \in \mathcal{F}$ such that $\mathcal{L}(\mathcal{T}_i) \cap A \neq \emptyset$ and $\mathcal{L}(\mathcal{T}_i) \cap (\overline{A} \cup \{\rho\}) \neq \emptyset$, and

(ii) for all $\mathcal{T}_i \in \mathcal{F}$, either $\mathcal{L}(\mathcal{T}_i) \subseteq A$ or $\mathcal{L}(\mathcal{T}_i) \subseteq (\overline{A} \cup \{\rho\})$.

Case (i). If there is a tree $\mathcal{T}_i$ with the properties in case (i), then $\mathcal{T}(\mathcal{L}_i)$ contains the root of $\mathcal{T}|A$. It follows that $\mathcal{T}_i$ is unique with the properties from (i) for otherwise $\mathcal{F}$ is not an agreement forest.

Now let $x \in \mathcal{L}(\mathcal{T}_i) \cap A$ and let $\mathcal{T}_{i,\overline{A}}$ be the tree obtained from $\mathcal{T}_i|(\overline{A} \cup \{\rho\} \cup \{x\})$ by relabelling x as a. Also, let $\mathcal{T}_{i,A}$ be the tree obtained from $\mathcal{T}_i|A$ by adding ρ at the end of a pendant edge adjoined to the root of $\mathcal{T}_i|A$.

Then

$$\mathcal{F}_A = \{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq A\} \cup \{T_{i,A}\}$$

is a good agreement forest for $T|A$ and $T'|A$ and

$$\mathcal{F}_{\overline{A},a} = \{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq \overline{A} \cup \{\rho\}\} \cup \{\mathcal{T}_{i,\overline{A}}\}$$

is a good agreement forest for $\overline{T}$ and $\overline{T'}$. Hence, $|\mathcal{F}| = |\mathcal{F}_A| + |\mathcal{F}_{\overline{A},a}| - 1$, and

$$\begin{aligned} h(\mathcal{T}, \mathcal{T}') &= |\mathcal{F}| - 1 \\ &= |\mathcal{F}_A| + |\mathcal{F}_{\overline{A}}| - 1 - 1 \\ &\geq h(\mathcal{T}|A, \mathcal{T}'|A) + h(\overline{\mathcal{T}}, \overline{\mathcal{T}'}). \end{aligned}$$

Case (ii). Since $\mathcal{G}_{\mathcal{F}}$ does not contain directed cycles it follows that the subdigraph of $\mathcal{G}_{\mathcal{F}}$ corresponding to the set $\{\mathcal{T}_i \in \mathcal{F} : \mathcal{L}(\mathcal{T}_i) \subseteq A\}$ does not contain directed cycles. Thus, this subdigraph has a vertex of in-degree zero, say $\mathcal{T}_0$. Let $\mathcal{T}_{0,\rho}$ be the tree obtained from $\mathcal{T}_0$ by adding ρ at the end of a pendant edge adjoined to the root of $\mathcal{T}_0$. Then

$$\mathcal{F}_A = \{\mathcal{T}_i \in \mathcal{F} : \mathcal{L}(\mathcal{T}_i) \subseteq A\} - \{\mathcal{T}_0\} \cup \{\mathcal{T}_{0,\rho}\}$$

is a good agreement forest for $\mathcal{T}|A$ and $\mathcal{T}'|A$. Also,

$$\mathcal{F}_{\overline{A},a} = \{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq \overline{A} \cup \{\rho\}\} \cup \{a\}$$

is a good agreement forest for $\overline{\mathcal{T}}$ and $\overline{\mathcal{T}'}$. So, in this case, $|\mathcal{F}| = |\mathcal{F}_A| + |\mathcal{F}_{\overline{A},a}| - 1$, and

$$\begin{aligned} h(\mathcal{T},\mathcal{T}') &= |\mathcal{F}| - 1 \\ &= |\mathcal{F}_A| + |\mathcal{F}_{\overline{A}}| - 1 - 1 \\ &\geq h(\mathcal{T}|A,\mathcal{T}'|A) + h(\overline{\mathcal{T}},\overline{\mathcal{T}'}). \end{aligned}$$

This completes the proof of the proposition. □

Note that a similar result does not hold for d_{rSPR} as the following example shows [12]. Consider the trees $\mathcal{T}$ and $\mathcal{T}'$ in Figure 5.4. Then $A = \{1,2,3,4\}$ is a common cluster for $\mathcal{T}$ and $\mathcal{T}'$. It is easily seen that $d_{rSPR}(\mathcal{T},\mathcal{T}') = 3$ and $d_{rSPR}(\mathcal{T}|A,\mathcal{T}'|A) = 2$. Also, $d_{rSPR}(\overline{\mathcal{T}},\overline{\mathcal{T}'}) = 2$.

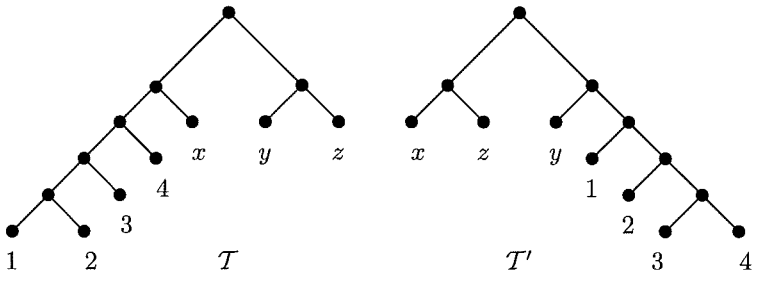


Figure 5.4: Two rooted phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ with a common cluster $A = \{1,2,3,4\}$ such that $d_{rSPR}(\mathcal{T}|A,\mathcal{T}'|A) + d_{rSPR}(\overline{\mathcal{T}},\overline{\mathcal{T}'}) > d_{rSPR}(\mathcal{T},\mathcal{T}')$.

If the collection of common clusters of $\mathcal{T}$ and $\mathcal{T}'$ partition X, the problem of finding $h(\mathcal{T},\mathcal{T}')$ is reduced to computing the minimum hybridization number for the restriction of the two trees to each common cluster.

Corollary 5.2.6. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Suppose that there exists a partition $\{A_1, A_2, \ldots, A_p\}$ of X such that $A_i \in c(\mathcal{T}) \cap c(\mathcal{T}')$ for all $i \in \{1,2,\ldots,p\}$. For each $i \in \{1,2,\ldots,p\}$ replace $\mathcal{T}|A_i$ respectively $\mathcal{T}'|A_i$ by a new leaf $l_i \notin X$, and denote by $\overline{\mathcal{T}}$ and $\overline{\mathcal{T}'}$ the trees obtained from $\mathcal{T}$ respectively $\mathcal{T}'$ in this way. Then $\mathcal{L}(\overline{\mathcal{T}}) = \mathcal{L}(\overline{\mathcal{T}'}) = \{l_1, l_2, \ldots l_p\}$ and*

$$h(\mathcal{T},\mathcal{T}') = h(\overline{\mathcal{T}},\overline{\mathcal{T}'}) + \sum_{i=1}^{p} h(\mathcal{T}|A_i,\mathcal{T}'|A_i).$$

Proof. The case $p = 1$ holds by Proposition 5.2.5. Assume now that $p \geq 2$ and construct the trees $\mathcal{T}_i$ and $\mathcal{T}'_i$, $i \in \{1, 2, \ldots, p\}$, as follows. Set $\mathcal{T}_0 = \mathcal{T}$ and $\mathcal{T}'_0 = \mathcal{T}'$. Having constructed the trees $\mathcal{T}_k$ and $\mathcal{T}'_k$, $0 \leq k < i$, let $\mathcal{T}_i$ be the rooted phylogenetic tree obtained from $\mathcal{T}_{i-1}$ by replacing $\mathcal{T}|A_i$ with the leaf l_i. Similarly, we obtain $\mathcal{T}'_i$ from $\mathcal{T}'_{i-1}$. Then for each $i \in \{1, 2, \ldots, p\}$, $\mathcal{T}_{i-1}|A_i = \mathcal{T}|A_i$ and $\mathcal{T}'_{i-1}|A_i = \mathcal{T}'|A_i$. Furthermore $\mathcal{T}_p = \overline{\mathcal{T}}$ and $\mathcal{T}'_p = \overline{\mathcal{T}'}$.

By applying Proposition 5.2.5 we obtain

$$h(\mathcal{T}_{i-1}, \mathcal{T}'_{i-1}) = h(\mathcal{T}_i, \mathcal{T}'_i) + h(\mathcal{T}|A_i, \mathcal{T}'|A_i), \ (1 \leq i \leq p),$$

and by summation we have

$$h(\mathcal{T}, \mathcal{T}') = h(\overline{\mathcal{T}}, \overline{\mathcal{T}'}) + \sum_{i=1}^{p} h(\mathcal{T}|A_i, \mathcal{T}'|A_i).$$

□

Corollary 5.2.7. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X–trees. Suppose that there exist $A_1, A_2 \in c(\mathcal{T}) \cap c(\mathcal{T}')$ such that $A_1 \cup A_2 = X$. Then*

$$h(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}|A_1, \mathcal{T}'|A_1) + h(\mathcal{T}|A_2, \mathcal{T}'|A_2).$$

Proof. By applying Corollary 5.2.6 we obtain that

$$h(\mathcal{T}, \mathcal{T}') = h(\overline{\mathcal{T}}, \overline{\mathcal{T}'}) + h(\mathcal{T}|A_1, \mathcal{T}'|A_1) + h(\mathcal{T}|A_2, \mathcal{T}'|A_2).$$

Then observe that $\overline{\mathcal{T}}$, $\overline{\mathcal{T}'}$ are isomorphic as they are rooted binary phylogenetic trees with the leaves labelled by l_1 and l_2, so $h(\overline{\mathcal{T}}, \overline{\mathcal{T}'}) = 0$. The result now follows. □

The result obtained in Corollary 5.2.6 suggests, for this particular case, a method to construct a minimal hybrid:

- Consider common clusters A_i, $1 \leq i \leq k$ of $\mathcal{T}$ and $\mathcal{T}'$ that partition X.
- In $\mathcal{T}$ (respectively $\mathcal{T}'$) replace each cluster A_i by a new leaf l_i. Denote by $\overline{\mathcal{T}}$ (respectively by $\overline{\mathcal{T}'}$) the trees obtained in this way. Construct a hybrid $\overline{\mathcal{H}}$ for these trees; $\mathcal{L}(\mathcal{H}) = \{l_1, l_2, \ldots, l_k\}$.
- For each i, construct a hybrid $\mathcal{H}_i$ corresponding to $\mathcal{T}|A_i$ and $\mathcal{T}'|A_i$.
- Replace each leaf l_i by $\mathcal{H}_i$ and obtain a hybrid $\mathcal{H}$.

5.3 An example from biology

Consider the two trees in Figure 5.6. Our aim is to estimate the minimum hybridization number of a hybrid that displays both trees. The trees $\mathcal{T}$ and $\mathcal{T}'$ are not binary. However, we denote by $h(\mathcal{T}, \mathcal{T}')$ the minimum number of hybrid events required to display $\mathcal{T}$ and $\mathcal{T}'$ by a single hybrid phylogeny and we show how we can restrict our discussion to the binary case. Then the main idea is to replace common clusters by leaves and apply Corollary 5.2.6. If for a cluster $C \in c(\mathcal{T}) \cap c(\mathcal{T}')$, $\mathcal{T}|C$ equals or is a refinement of $\mathcal{T}'|C$ (or vice versa) then we will assume that both can be replaced by the same binary tree and consequently $h(\mathcal{T}|C, \mathcal{T}'|C) = 0$. The common clusters we consider and the relationship between the restriction of $\mathcal{T}$ (respectively $\mathcal{T}'$) to each cluster are represented in Figure 5.5. Note that $\{A_1, A_2, \ldots, A_{15}\}$ is a partition of X and each A_i is a common cluster of the trees $\mathcal{T}$ and $\mathcal{T}'$.

A_1	$1, 2, 3, 4, 5$	$\mathcal{T}'\|A_1$ is a refinement of $\mathcal{T}\|A_1$
A_2	$6, 7, \ldots, 13$	$\mathcal{T}\|A_2$ is a refinement of $\mathcal{T}'\|A_2$
A_3	$17, 18, 19, 20$	$\mathcal{T}\|A_3$ is a refinement of $\mathcal{T}'\|A_3$
A_4	$15, 16$	$\mathcal{T}\|A_4 = \mathcal{T}'\|A_4$
A_5	14	$\mathcal{T}\|A_5 = \mathcal{T}'\|A_5$
A_6	$21, 22, \ldots, 33$	$2 \leq h(\mathcal{T}\|A_6, \mathcal{T}'\|A_6) \leq 3$
A_7	$34, 35$	
A_8	$36, 37$	
A_9	38	$\mathcal{T}\|A_i = \mathcal{T}'\|A_i$, $7 \leq i \leq 13$
A_{10}	39	
A_{11}	40	
A_{12}	41	
A_{13}	42	
A_{14}	$43, 44, 45$	$\mathcal{T}\|A_{14}$ is a refinement of $\mathcal{T}'\|A_{14}$
A_{15}	46	$\mathcal{T}\|A_{14} = \mathcal{T}'\|A_{14}$

Figure 5.5: The clusters we consider for the trees in Figure 5.6.

Then let us observe that by pruning from $\mathcal{T}$, respectively from $\mathcal{T}'$, the subtrees

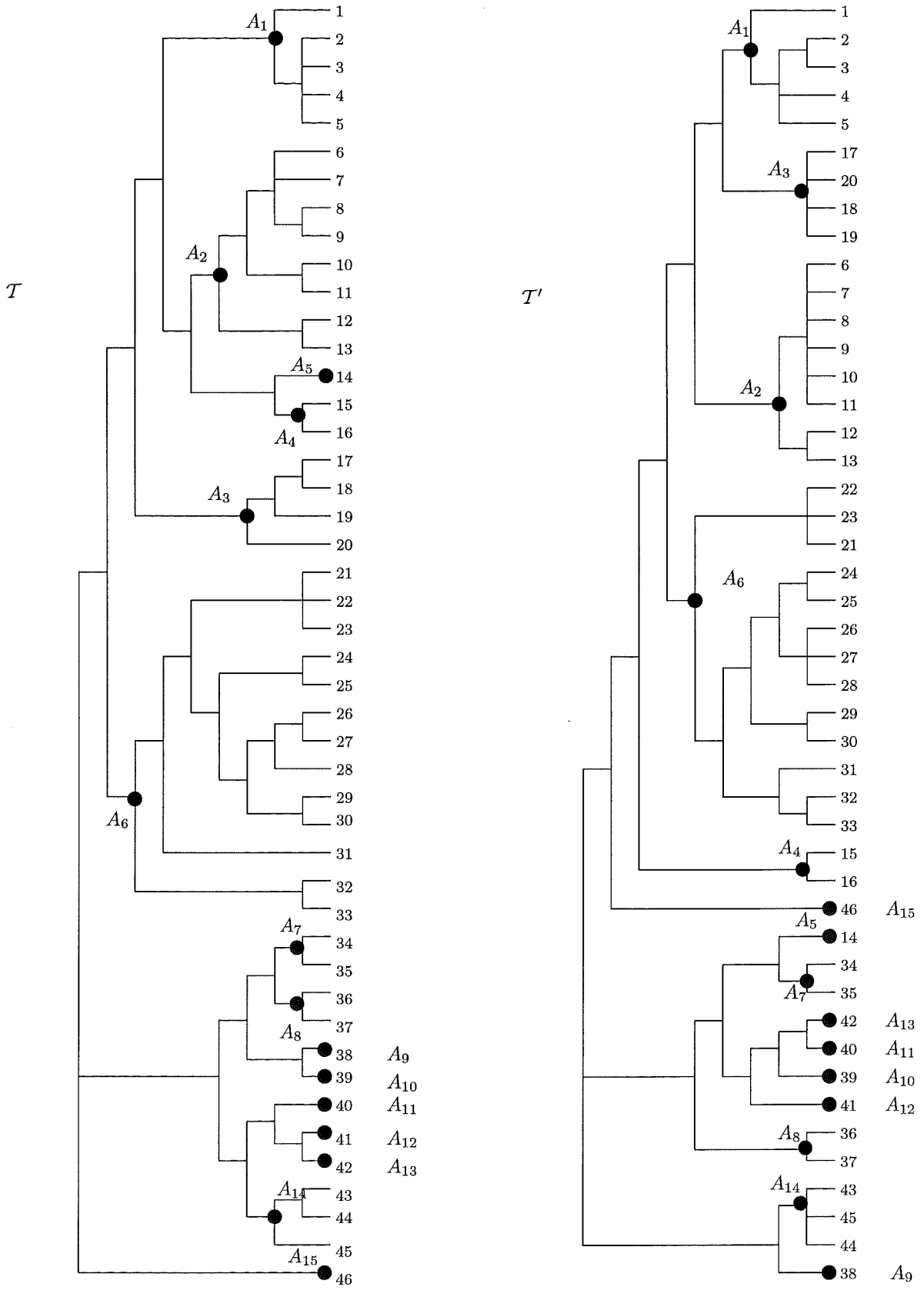


Figure 5.6: Alpine Ranunculi of New Zealand, Quartet puzzle phylogenetic trees for nuclear ITS sequence ($\mathcal{T}$), respectively, J_{SA} sequence ($\mathcal{T}'$). Adapted from [34].

corresponding to clusters A_9, A_{14}, and A_{15} we obtain two binary trees $\mathcal{T}_1$ and $\mathcal{T}_1'$ and, by applying Proposition 5.2.1 and Corollary 5.2.3 we have that:

$$h(\mathcal{T}, \mathcal{T}') - 3 \leq h(\mathcal{T}_1, \mathcal{T}_1') \leq h(\mathcal{T}, \mathcal{T}').$$

Let $\overline{\mathcal{T}_1}$, $\overline{\mathcal{T}_1'}$ be the trees obtained from $\mathcal{T}_1$ (respectively $\mathcal{T}_1'$) by replacing $\mathcal{T}_1|A_i$ and $\mathcal{T}_1'|A_i$ by the same leaf l_i. Denote by $\mathcal{L}_1 = \mathcal{L}(\overline{\mathcal{T}_1}) = \mathcal{L}(\overline{\mathcal{T}_1'})$. If we restrict $\overline{\mathcal{T}_1}$ (respectively $\overline{\mathcal{T}_1'}$) to $\mathcal{L}_2 = \mathcal{L}_1 - \{l_5\}$ we simplify the problem as the obtained trees , $\overline{\mathcal{T}_2}$ and $\overline{\mathcal{T}_2'}$ say, have two common clusters $C_1 = \{l_1, l_2, l_3, l_4, l_6\}$ and $C_2 = \{l_7, l_8, l_{10}, l_{11}, l_{12}, l_{13}\}$. On the other hand, from Corollary 5.2.2 it follows that

$$h(\overline{\mathcal{T}_1}, \overline{\mathcal{T}_1'}) - 1 \leq h(\overline{\mathcal{T}_2}, \overline{\mathcal{T}_2'}) \leq h(\overline{\mathcal{T}_1}, \overline{\mathcal{T}_1'}).$$

These trees are drawn in Figure 5.7.

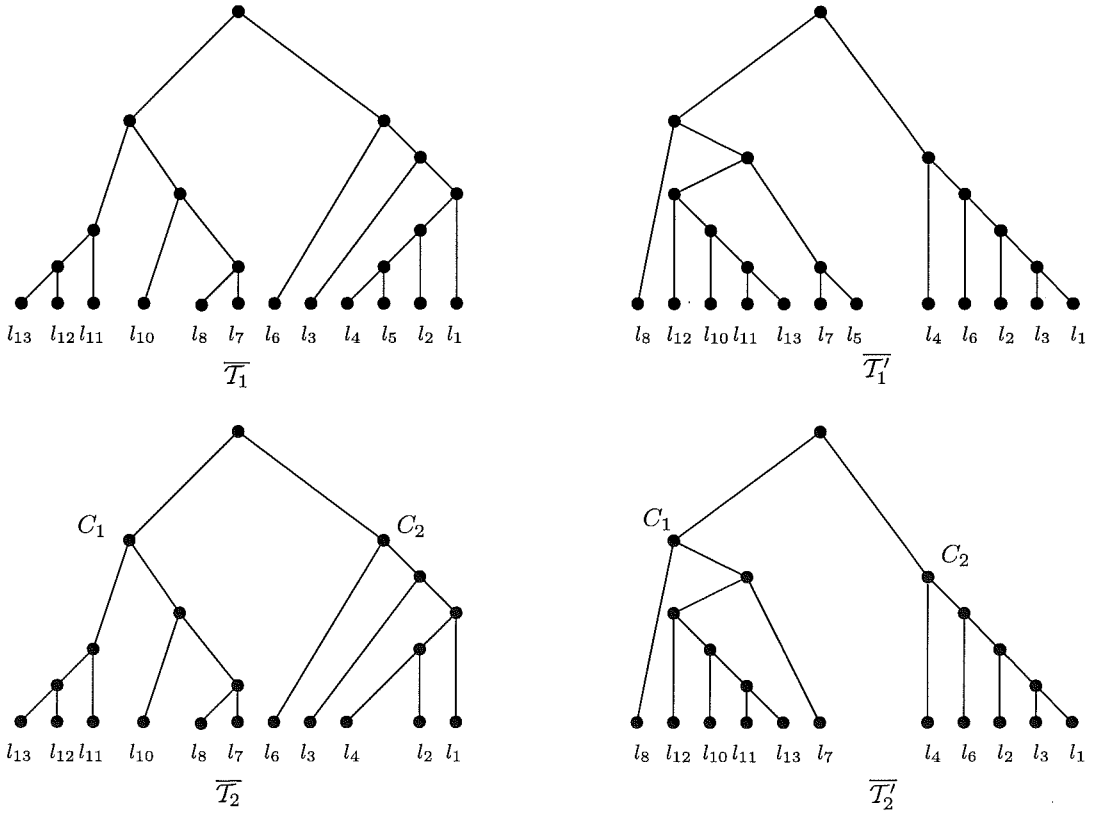


Figure 5.7:

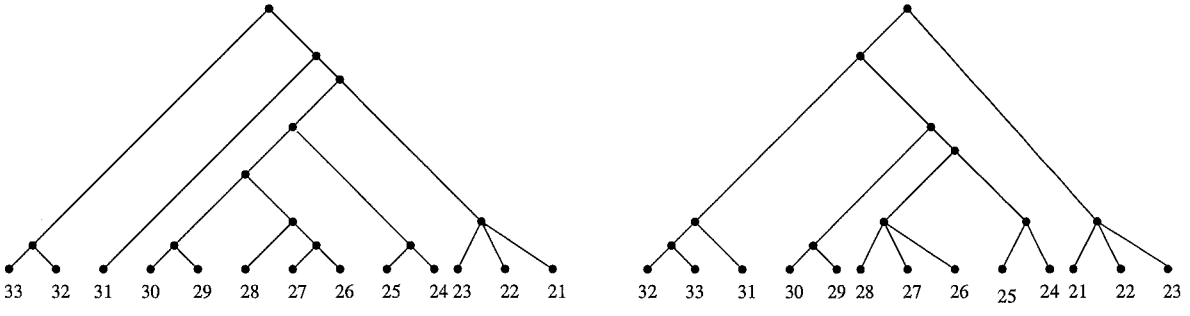
Now for $\overline{T_2}$ and $\overline{T_2'}$ and the clusters C_1 and C_2 we can apply Corollary 5.2.7. It follows that

$$h(\overline{T_2}, \overline{T_2'}) = h(\overline{T_2}|C_1, \overline{T_2'}|C_1) + h(\overline{T_2}|C_2, \overline{T_2'}|C_2) = 2 + 3 = 5.$$

Therefore, we obtain that

$$5 \leq h(\overline{T_1}, \overline{T_1'}) \leq 6.$$

Using the same ideas, it is straigtforward to show that $2 \leq h(\mathcal{T}|A_6, \mathcal{T}'|A_6) \leq 3$ (see Figure 5.8).



$\mathcal{T}|A_6$ $\mathcal{T}'|A_6$

Figure 5.8:

Taking into account the previous observations and Proposition 5.2.5, we have

$$h(\mathcal{T}_1, \mathcal{T}_1') = h(\overline{T_1}, \overline{T_1'}) + h(\mathcal{T}|A_6, \mathcal{T}'|A_6),$$

and thus

$$7 \leq h(\mathcal{T}_1, \mathcal{T}_1') \leq 9.$$

Finally, we obtain

$$7 \leq h(\mathcal{T}, \mathcal{T}') \leq 12.$$

Now let us consider the restrictions t and t' of $\mathcal{T}$ (respectively $\mathcal{T}'$) to the set $\mathcal{L} = \{1, 2, \ldots, 20\}$ (this corresponds to the clusters A_i, $1 \leq i \leq 5$ in Figure 5.6), and apply the construction described in the end of Section 5.2 to construct a minimal hybrid that displays t and t'. For each $i = 1, 5$ replace A_i by x_i. The corresponding

trees $\overline{t}$ and $\overline{t'}$ are shown in Figure 5.9 a). Note that, in this case, $d_{rSPR}(\overline{t}, \overline{t'}) = h(\overline{t}, \overline{t'}) = 3$.

Two minimal regular hybrids, $\mathcal{H}_1$ and $\mathcal{H}_2$, that displays $\overline{t}$ and $\overline{t'}$ are drawn in Figure 5.9 b). Note that $\mathcal{H}_1$ is regular, but for the regularity of $\mathcal{H}_2$, two extra leaves, y_1 and y_2, have been added.

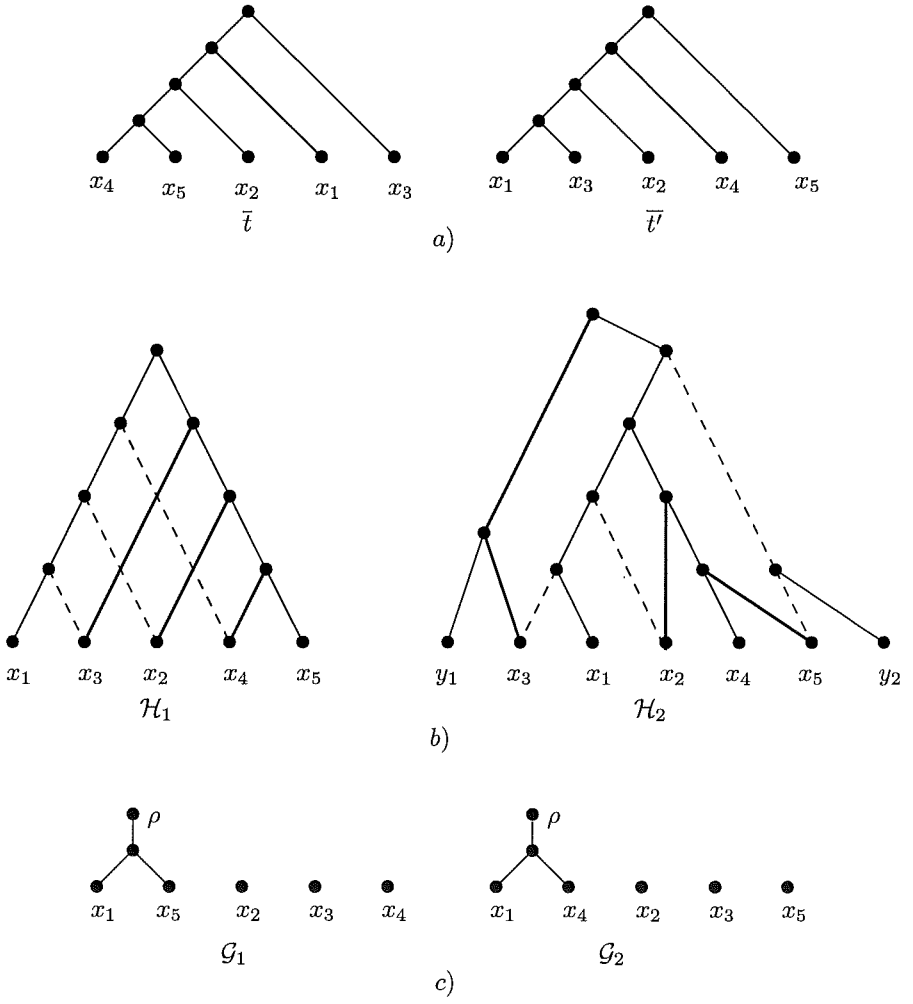


Figure 5.9: The hybrids $\mathcal{H}_1$ and $\mathcal{H}_2$ correspond to two sequences of rSPR operations that transform $\overline{t'}$ into $\overline{t}$. The associated agreement forests are $\mathcal{G}_1$, and $\mathcal{G}_2$ respectively.

Chapter 6

Accumulation phylogenies

In this chapter, we formalize and analyse a simple model in which evolved characteristics are passed on to all descendant species. We obtain two main results. First we show that the resulting observed sets of characteristics for the species at the leaves uniquely determine the digraph that described the evolution of the species, under certain restrictions. Second, we characterize when this digraph is actually a tree.

6.1 Accumulation phylogenies

Let S be any (finite or infinite) set. An *accumulation phylogeny on* X is a triple (V, A, f) where (V, A) is a hybrid phylogeny on X with the root ρ, and f is a function from $A \cup \{\rho\}$ to 2^S (the power set of S) that satisfies the following two properties:

(P1) the set $\{f(a) : a \in A \cup \{\rho\}\}$ is a collection of pairwise disjoint sets; and

(P2) for each vertex $v \in V - \{\rho\}$ we have $f(a) \neq \emptyset$ for at least one arc a that ends at v.

Note that $f(\rho)$, and $f(a)$ (for certain arcs a) can be the empty set.

In molecular evolutionary biology we may view (V, A) as modelling the evolution of a collection X of extant species from a common ancestor ρ, allowing for hybrid or reticulate evolution. In this setting we may regard S as a set of genes, $f(\rho)$ as the genes present in the most recent common ancestor of the species in X, and for each arc $a = (u, v)$, $f(a)$ denotes the genes that arose in the lineage leading from (ancestral species) u to v. An accumulation phylogeny models the situation where (i) genes arise at most once (condition (**P1**)), a situation that is commonly assumed and goes under the name 'gene genesis'; and (ii) if a new gene is subsequently lost, some 'trace' of that gene is at least detectable in all descendant species. Furthermore, we assume that S is sufficiently extensive that each descendant species acquires at least one new gene (condition (**P2**)).

Recently, the gene content of species has been used for reconstructing phylogenies (see [54]) by constructing measures of (dis)similarity based on the amount of genes shared by two species (we consider this further in the next section). The approach we consider here may provide an alternative technique for reconstructing the phylogenetic history of species (not necessarily based on a phylogenetic tree) using gene content or other types of genomic markers.

6.2 Accumulation maps

Suppose that (V, A, f) is an accumulation phylogeny on X. Consider the map $\overline{\alpha} : V \to 2^S$, defined by setting

$$\overline{\alpha}(v) := f(\rho) \cup \bigcup_{a \in \text{path from } \rho \text{ to } v} f(a).$$

Let $\alpha := \overline{\alpha}|X$, the restriction of the map $\overline{\alpha}$ to X. We refer to α (respectively $\overline{\alpha}$) as the **accumulation map** on X (respectively on V) **induced** by (V, A, f) (or, more briefly, by (V, A)).

Informally, the map $\overline{\alpha}$ describes how the elements of S 'accumulate' as one moves down the digraph, and $\alpha(x)$ describes the resulting subset of S at the leaf x. For example, for either of the hybrid phylogenies shown in Figure 6.1 together with

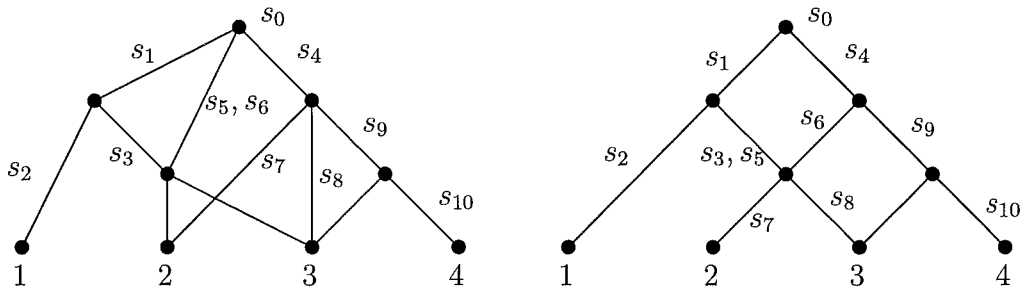


Figure 6.1: Two accumulation phylogenies on $X = \{1, 2, 3, 4\}$ with the same induced accumulation map α.

the values $f(\rho)$ and $f(a)$ as indicated (from the set $\{s_0, s_1, \ldots s_{10}\}$), one has $\alpha(2) = \{s_0, s_1, s_3, s_4, s_5, s_6, s_7\}$. Our interest in this thesis is in studying what the $\alpha(x)$ values tell us about the underlying accumulation phylogeny (V, A, f). As Figure 6.1 shows, different accumulation phylogenies can give rise to the same accumulation map on X.

6.3 Properties of α and $\overline{\alpha}$

We now provide three lemmas that describe the basic properties of accumulation maps. These results will be useful later in the chapter. First, we define some terminology.

Consider a map $\alpha : X \to 2^S$. For $s \in S$, let

$$A(s) := \{x \in X : s \in \alpha(x)\},$$

and consider the associated family of subsets of X:

$$\mathcal{A}(\alpha) := \{A(s) : s \in S\} \cup \{X\}.$$

Lemma 6.3.1. *Suppose that (V, A, f) is an accumulation phylogeny on X, with induced accumulation map α, and with $S = \cup_{x \in X} \alpha(x)$.*

(i) For each arc $a = (u, v) \in A$, and each $s \in f(a)$ we have $A(s) = c(v)$.

(ii) *For each* $s \in S$, $A(s) = c(v)$ *for at least one vertex* $v \in V$.

(iii) *For each vertex* $v \in V - \{\rho\}$ *there exists at least one element* $s \in S$ *such that* $c(v) = A(s)$.

(iv) $\mathcal{A}(\alpha) = \{c(v) : v \in V\}$.

Proof. (i) First suppose that there exists $a \in A$ with $s \in f(a)$, and $a = (u, v)$ for some $u \in V$. If $x \in c(v)$ then $v \leq_{\mathcal{D}} x$ and so $s \in \alpha(x)$. Consequently, $c(v) \subseteq A(s)$. To establish the reverse inclusion, suppose that $x \in A(s)$. Then by **(P1)** there exists a path from v to x and so $x \in c(v)$. Hence $A(s) \subseteq c(v)$, and thus $A(s) = c(v)$ as claimed.

(ii) Suppose that $s \in S$. Since $S = \cup_{x \in X} \alpha(x)$, there exists an element $w \in A \cup \{\rho\}$ with $s \in f(w)$. By **(P1)** this element w is unique. If $w = \rho$ then $A(s) = X = c(\rho)$. Otherwise, if w is an arc, let v denote the end vertex of w. By Part (i) we have that $A(s) = c(v)$.

(iii) For any vertex $v \in V - \{\rho\}$, **(P2)** guarantees that there is at least one arc a that ends at v for which $f(a) \neq \emptyset$. Select any $s \in f(a)$. Then $c(v) = A(s)$ by Part (i).

(iv) Let $A \in \mathcal{A}(\alpha)$. Then $A = X = c(\rho)$ or there exists $s \in S$ with $A = A(s)$. In the latter case, from (ii), it follows that $A(s) = c(v)$ for some $v \in V$. Conversely, let $v \in V$. If $v = \rho$, then $c(v) = X \in \mathcal{A}(\alpha)$. If $v \neq \rho$, from (iii), it follows that there exists $s \in S$ such that $c(v) = A(s)$, and therefore $c(v) \in \mathcal{A}(\alpha)$. □

Notice that, by Lemma 6.3.1, part (iv), $\mathcal{A}(\alpha)$ does not depend on f.

Lemma 6.3.2. *Let* (V, A, f) *be an accumulation phylogeny on* X, *and let* α *and* $\overline{\alpha}$ *be the induced accumulation maps on* X *and* V *respectively. Let* $\mathcal{D} = (V, A)$.

(1) *For any* $u, v \in V$, *the following conditions are satisfied:*

$$(i)\ u \neq v \Leftrightarrow \overline{\alpha}(u) \neq \overline{\alpha}(v);$$
$$(ii)\ u <_{\mathcal{D}} v \Leftrightarrow \overline{\alpha}(u) \subset \overline{\alpha}(v);$$
$$(iii)\ u \leq_{\mathcal{D}} v \Leftrightarrow \overline{\alpha}(u) \subseteq \overline{\alpha}(v).$$

(2) For any $v \in V$,

$$\overline{\alpha}(v) \subseteq \bigcap_{x \in c(v)} \alpha(x).$$

(3) Given $v \in V$,

$$\bigcap_{x \in c(v)} \alpha(x) \subseteq \overline{\alpha}(v) \Leftrightarrow (c(v) \subseteq c(u) \Rightarrow u \leq_{\mathcal{D}} v).$$

Proof. (1)–(i) Suppose $u \neq v$. The conditions $u <_{\mathcal{D}} v$ and $v <_{\mathcal{D}} u$ cannot be satisfied simultaneously. Assuming that $u <_{\mathcal{D}} v$ does not hold, it follows that $u \neq \rho$ and any arc ending in u does not lie on a path from ρ to v. Let a be such an arc with $f(a) \neq \emptyset$. Then $f(a) \subseteq \overline{\alpha}(u) - \overline{\alpha}(v)$, hence $\overline{\alpha}(u) \neq \overline{\alpha}(v)$.

(1)–(ii) Assume that $u <_{\mathcal{D}} v$ and let $s \in \overline{\alpha}(u)$. Then either $s \in f(\rho)$ or $s \in f(a)$ for some arc a in a path from ρ to u. In the former case $s \in \overline{\alpha}(v)$ and in the latter a is an arc of a path from ρ to v. It follows from Part 1(i) that the inclusion is strict.

Conversely, suppose that $\overline{\alpha}(u) \subset \overline{\alpha}(v)$. If $u = \rho$, then $u <_{\mathcal{D}} v$. Assume now that $u \in V - \{\rho\}$. According to **(P2)** there exists s in S such that $s \in f(a)$ for some arc a ending in u. Therefore $s \in \overline{\alpha}(u)$, and hence $s \in \overline{\alpha}(v)$. From **(P1)** it follows that $s \in f(b)$ entails $b = a$. Since $s \in \overline{\alpha}(v)$, there is a path from ρ to v containing a. Consequently, $u <_{\mathcal{D}} v$.

(1)–(iii) This follows from parts 1(i) and 1(ii).

(2) Let $x \in c(v)$. Since $v \leq_{\mathcal{D}} x$, we have $\overline{\alpha}(v) \subseteq \alpha(x)$. Consequently, $\overline{\alpha}(v) \subseteq \cap_{x \in c(v)} \alpha(x)$.

(3) Let $v \in V$ and suppose that $\cap_{x \in c(v)} \alpha(x) \subseteq \overline{\alpha}(v)$. Let u be a vertex of V with $c(v) \subseteq c(u)$. It follows that

$$\overline{\alpha}(u) \subseteq \bigcap_{x \in c(u)} \alpha(x) \subseteq \bigcap_{x \in c(v)} \alpha(x) \subseteq \overline{\alpha}(v),$$

and so by part 1–(iii) of this lemma, $u \leq_{\mathcal{D}} v$.

Conversely, let $v \in V$ and suppose that $c(v) \subseteq c(u)$ entails $u \leq_{\mathcal{D}} v$. Let $s \in \cap_{x \in c(v)} \alpha(x)$. If $s \in f(\rho)$ then $s \in \overline{\alpha}(v)$. Otherwise, $s \in f(a)$ for some arc $a = (w, u)$.

For any $x \in c(v)$ there is at least one path from ρ to x, containing a. It follows that $c(v) \subseteq c(u)$, and thus $u \leq_{\mathcal{D}} v$. Therefore, $s \in \overline{\alpha}(v)$. □

Given a rooted digraph (V, A) and any vertex v of V let $end(v)$ be the set

$$end(v) := \begin{cases} \{\rho\}, \text{ if } v = \rho; \\ \{(u, v) : (u, v) \in A\}, \text{ otherwise;} \end{cases}$$

and for any function $f : A \cup \{\rho\} \to 2^S$ let

$$f(end(v)) := \{s \in S : s \in f(a) \text{ for some } a \in end(v)\}.$$

Lemma 6.3.3. *Let (V, A, f) be an accumulation phylogeny on X, and let $\overline{\alpha}$ be the associated accumulation map on V. For all vertices v of V,*

$$f(end(v)) = \overline{\alpha}(v) - \bigcup_{(u,v)\in A} \overline{\alpha}(u).$$

Proof. Let $v \in V$. If $v = \rho$ then $f(end(v)) = f(\rho) = \overline{\alpha}(\rho)$ and there is no arc (u, v) in A. If $v \neq \rho$, then

$$\overline{\alpha}(v) = \left(\bigcup_{(u,v)\in A} \overline{\alpha}(u) \right) \cup f(end(v))$$

and the two sets in the union are disjoint. The lemma now follows. □

6.4 Regular hybrid phylogenies

For accumulation phylogenies for which the underlying digraph $\mathcal{D}$ is regular, the induced map α suffices (along with $\mathcal{D}$) to reconstruct the sets $\overline{\alpha}(v)$ and $f(end(v))$ for every vertex v of V, as we now show.

Proposition 6.4.1. *Suppose $\mathcal{D} = (V, A)$ is a regular hybrid phylogeny, and that α is an accumulation map induced by (V, A, f). Then for any $v \in V$,*

(i)

$$\overline{\alpha}(v) = \bigcap_{x\in c(v)} \alpha(x),$$

(ii)

$$f(end(v)) = \bigcap_{x \in c(v)} \alpha(x) - \bigcup_{(u,v) \in A} \bigcap_{x \in c(u)} \alpha(x).$$

Proof. (i) We have $\overline{\alpha}(v) \subseteq \bigcap_{x \in c(v)} \alpha(x)$, by Lemma 6.3.2(2). The reverse inclusion follows from Lemma 6.3.2(3), since if (V, A) is regular then for $u, v \in V$ we have $c(v) \subseteq c(u)$ entails $u \leq_{\mathcal{D}} v$.

(ii) By Lemma 6.3.3 we may identify $f(end(v))$ with $\overline{\alpha}(v) - \bigcup_{(u,v) \in A} \overline{\alpha}(u)$. By Part (i) of Proposition 6.4.1 we have $\overline{\alpha}(v) = \bigcap_{x \in c(v)} \alpha(x)$ and $\overline{\alpha}(u) = \bigcap_{x \in c(u)} \alpha(x)$ from which Part (ii) now follows. □

6.5 Representations of accumulation maps on X

Our first main theorem considers the existence and uniqueness of representations of an arbitrary map α by an accumulation phylogeny. The existence question has a fairly straightforward solution, but the uniqueness question is more interesting, because as Figure 6.1 showed, α does not, in general, uniquely determine the underlying hybrid phylogeny (V, A). However the following result shows that if we restrict attention to regular hybrid phylogenies then one can uniquely recover (V, A) (along with the sets $f(end(v))$) from α.

Theorem 6.5.1. *Let S be an arbitrary set, X a finite non-empty set, and α a map from X to 2^S with $S = \bigcup_{x \in X} \alpha(x)$. Then, α is the accumulation map induced by at least one accumulation phylogeny if and only if*

$$\alpha(x) - \bigcup_{y \in X : y \neq x} \alpha(y) \neq \emptyset, \text{ for all } x \in X. \tag{6.1}$$

Moreover, when this holds, there is a unique regular hybrid phylogeny (V, A) on X such that α is the accumulation map induced by (V, A, f) for at least one choice of a map f. Although f is not necessarily uniquely determined by α the sets $f(end(v))$, $v \in V$ are uniquely determined by α.

Proof. If α is the accumulation map on X for an accumulation phylogeny (V, A, f) then inequality (6.1) follows from the property (**P2**).

Assume now that (6.1) holds. For each $s \in S$, let $A(s) = \{x \in X : s \in \alpha(x)\}$ and let $x \in X$. It follows that $A(s) = \{x\}$ for each $s \in \alpha(x) - \cup_{y \neq x} \alpha(y)$. Let $\mathcal{C} = \{A(s) : s \in S\} \cup \{X\}$ and let $\mathcal{D} = (V, A)$ be the cover digraph of $\mathcal{C}$. Note that $\mathcal{D}$ is a regular hybrid phylogeny. Then define $\overline{\alpha} : V \to 2^S$, by setting

$$\overline{\alpha}(v) = \bigcap_{x \in c(v)} \alpha(x),$$

and for each $v \in V$, let

$$S_v = \overline{\alpha}(v) - \bigcup_{(u,v) \in A} \overline{\alpha}(u).$$

It suffices to construct a map f from $A \cup \{\rho\}$ to 2^S such that $f(end(v)) = S_v$ for each $v \in V$. To this end, for each $v \in V - \{\rho\}$, let

$$g_v : S_v \to end(v)$$

be an arbitrary function. Now define $f : A \cup \{\rho\} \to 2^S$, by setting $f(\rho) = \overline{\alpha}(\rho)$, and $f((u,v)) = \{s : g_v(s) = (u,v)\}$. Clearly, (V, A, f) is an accumulation phylogeny on X and $\overline{\alpha}$ is its induced accumulation map; furthermore $\alpha = \overline{\alpha}|X$. This establishes the first part of Theorem 6.5.1. The uniqueness of (V, A) follows from Lemma 6.3.1(iv) since we are assuming (V, A) is regular, and so it is determined by $\{c(v) : v \in V\}$. The uniqueness of $f(end(v))$ follows from Proposition 6.4.1(ii). □

Remarks

1. Referring to Theorem 6.5.1, since (V, A) is uniquely determined by α, so are the numbers $d^-(v)$. There are then exactly

$$\prod_{v \neq \rho} (d^-(v))^{|S_v|}$$

ways to define the map $f : A \cup \{\rho\} \to 2^S$. Consequently, f (and thereby the accumulation phylogeny (V, A, f)) is uniquely determined by α if and only if (V, A) is a tree.

2. Note that the unique regular hybrid phylogeny (V, A) that furnishes a representation of α can be reconstructed from α by an algorithm that runs in polynomial time (in $|X|$). Similarly, constructing a function f such that α is the accumulation map for (V, A, f) is computationally easy.

6.6 Recognizing trees

Accumulation maps that are induced by rooted phylogenetic trees give rise to a particular numerical and set-theoretic property. In this section we show how the set-theoretic (but not the numerical) property characterizes when the underlying digraph in an accumulation phylogeny is a rooted phylogenetic tree. We begin with some terminology.

For any map $\alpha : X \to 2^S$ define

$$D_\alpha : X \times X \to 2^S$$

by $D_\alpha(x, y) = \alpha(x) \cap \alpha(y)$.

Recall that an *ultrametric* on X is any symmetric function $d : X \times X \to \mathbf{R}$ that satisfies the following two properties:

- $d(x, x) = 0$ for all $x \in X$, and
- for any three distinct elements $x, y, z \in X$, $d(x, y) \leq \max\{d(x, z), d(y, z)\}$.

An example of an ultrametric on X is provided by any rooted phylogenetic tree $\mathcal{T}$ on X together with any function h from the vertices of $\mathcal{T}$ into the set of real numbers that is increasing along any path from a leaf to the root and setting $d(x, y) = h(\text{mrca}(x, y))$ for all distinct pairs x, y (and setting $d(x, x) = 0$ for all $x \in X$). In this case we say that $\mathcal{T}$ provides a *representation of* d. It is a classic result that for any ultrametric d there is a unique rooted phylogenetic tree on X that provides a representation of d (see eg. [48]).

If $\mathcal{D}$ is a phylogenetic tree then it is easy to recover $\mathcal{D}$ from any induced accumulation map on X by virtue of the following result.

Proposition 6.6.1. *Suppose that α is an accumulation map on X induced by a rooted phylogenetic tree $\mathcal{T}$. Define $d_\alpha : X \times X \to \mathbf{R}$ by setting*

$$d_\alpha(x,y) = \begin{cases} 0, & \text{if } x = y; \\ -|D_\alpha(x,y)|, & \text{otherwise.} \end{cases}$$

Then d_α is an ultrametric on X and $\mathcal{T}$ is the (unique) rooted phylogenetic tree on X that provides a representation of d_α.

A natural question at this point is whether the converse of Proposition 6.6.1 holds - that is, if d_α is an ultrametric for an accumulation map α induced by a hybrid phylogeny $\mathcal{D}$ does this imply that $\mathcal{D}$ is a rooted phylogenetic tree? The following example shows that the answer is no.

Example. Let $X = \{1,2,3\}$ and consider the cover digraph $\mathcal{D}$ of the following subsets: $\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}$; this digraph is shown in Figure 6.2. For the associated function f take $S = A$ and assign the singleton set $\{a\}$ to each arc a of $\mathcal{D}$. Then d_α is an ultrametric yet $\mathcal{D}$ is not a rooted phylogenetic tree.

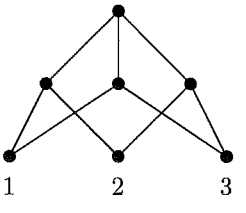


Figure 6.2:

This example shows that it is impossible to distinguish between a tree and a non-tree hybrid phylogeny solely on the basis of the number of elements of S that each pair of elements x and y differ on. This raises a further question of whether it is possible to recognize when the underlying digraph is a tree if one uses more of the structure of the set system $\{D_\alpha(x,y) : x,y \in X, x \neq y\}$ than just the cardinality of these sets.

To this end we say that D_α satisfies the **set–ultrametric property** if for any three distinct elements $x, y, z \in X$, two of the sets $D_\alpha(x, y)$, $D_\alpha(y, z)$ and $D_\alpha(x, z)$ are equal.

If in addition, the two equal sets are always contained (resp. strictly contained) in the third we say that D_α has the **nested** (respectively **strictly–nested**) **set–ultrametric property**.

We are now ready to state our second main result.

Theorem 6.6.2. *Let α be the accumulation map on X induced by a regular hybrid phylogeny $\mathcal{D} = (V, A)$.*

(1) The following are equivalent:

(i) $\mathcal{D} = (V, A)$ is a rooted phylogenetic tree,

(ii) D_α has the set–ultrametric property,

(iii) D_α has the nested set–ultrametric property.

(2) $\mathcal{D} = (V, A)$ is a rooted binary phylogenetic tree if and only if D_α has the strictly–nested set-ultrametric property.

Proof. (1) We first show that (i) implies (iii). Assume that $\mathcal{D}$ is a rooted phylogenetic tree on X and let $x, y, z \in X$. On the one hand, for any vertices u and v of V, $\overline{\alpha}(u) \cap \overline{\alpha}(v) = \overline{\alpha}(\mathrm{mrca}(u, v))$, where $\mathrm{mrca}(u, v)$ denotes the most recent common ancestor of u and v in $\mathcal{D}$. On the other hand, two of the vertices $\mathrm{mrca}(x, y)$, $\mathrm{mrca}(y, z)$, and $\mathrm{mrca}(z, x)$ are equal. We may assume that $\mathrm{mrca}(x, y) = \mathrm{mrca}(y, z)$. In this case, $\mathrm{mrca}(x, y) \leq_{\mathcal{D}} \mathrm{mrca}(x, z)$, and hence $\overline{\alpha}(\mathrm{mrca}(x, y)) = \overline{\alpha}(\mathrm{mrca}(y, z)) \subseteq \overline{\alpha}(\mathrm{mrca}(x, z))$. Therefore, since $D_\alpha(x_1, x_2) = \overline{\alpha}(\mathrm{mrca}(x_1, x_2))$, for any pair $x_1, x_2 \in X$ we have $\mathcal{D}_\alpha(x, y) = \mathcal{D}_\alpha(y, z) \subseteq \mathcal{D}_\alpha(x, z)$.

Condition (iii) clearly implies condition (ii). We now show that (ii) implies (i) – or equivalently, that the negation of (i) implies the negation of (ii). Thus let us assume that $\mathcal{D}$ is not a tree, then there exists a vertex v with $d^-(v) \geq 2$. Let u_1 and u_2 be two vertices such that $(u_1, v) \in A$, $(u_2, v) \in A$ and $u_1 \neq u_2$ (Figure 6.3). Since

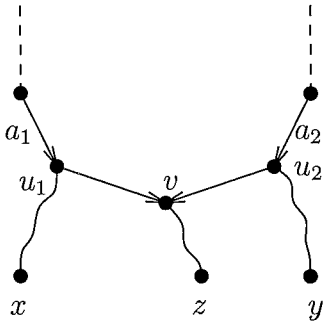


Figure 6.3:

$\mathcal{D}$ is regular there is no path from u_1 to u_2 or from u_2 to u_1. Consequently, neither of the inclusions $c(u_1) \subseteq c(u_2)$ and $c(u_2) \subseteq c(u_1)$ hold, so there exist two leaves x and y with $x \in c(u_1) - c(u_2)$ and $y \in c(u_2) - c(u_1)$. Let $z \in c(v)$.

The hybrid $\mathcal{D}$ is regular, so $u_1 \neq \rho \neq u_2$, and hence, $d^-(u_1) \neq 0 \neq d^-(u_2)$. For each $i \in \{1, 2\}$ let a_i be an arc ending in u_i with $f(a_i) \neq \emptyset$. Then, $f(a_1) \subseteq \alpha(x) \cap \alpha(z)$, $f(a_2) \subseteq \alpha(y) \cap \alpha(z)$, $f(a_1) \cap \alpha(y) = \emptyset = f(a_2) \cap \alpha(x)$. Therefore, the sets $\mathcal{D}_\alpha(x, y)$, $\mathcal{D}_\alpha(y, z)$, and $\mathcal{D}_\alpha(x, z)$ are mutually distinct, which contradicts the set–ultrametric property.

(2) If, in addition, $\mathcal{D}$ is binary and $x, y, z \in X$ are mutually distinct, then $\mathrm{mrca}(x, y) \neq \mathrm{mrca}(x, z)$, hence the containment is strict. Conversely, assume that the strictly-nested set-ultrametric property is satisfied. According to (i), $\mathcal{D} = (V, A)$ is a tree. If $\mathcal{D}$ is not binary one can find the vertices u, v_1, v_2, v_3 such that $(u, v_i) \in A$, $1 \leq i \leq 3$. If $x_i \in c(v_i)$, then $D_\alpha(x_1, x_2) = D_\alpha(x_2, x_3) = D_\alpha(x_3, x_1) = \overline{\alpha}(u)$, which is contradictory to the strictly-nested set-ultrametric property. □

Let M be an arbitrary set. The map δ from $X \times X$ into M is said to be an *symbolic ultrametric* (*on* X) if each of the following conditions are satisfied:

(U1) $\delta(i, j) = \delta(j, i)$, for all $i, j \in X$;

(U2) $|\{\delta(i, j), \delta(i, k), \delta(j, k)\}| \leq 2$, for all $i, j, k \in X$;

(U3) there are no pairwise distinct elements i, j, k, and l of X with

$$\delta(i,j) = \delta(j,k) = \delta(k,l) \neq \delta(j,l) = \delta(l,i) = \delta(i,k).$$

Any ultrametric on X is a symbolic ultrametric on X.

It is easily seen that, for any map $\alpha : X \to 2^S$, the associated function D_α satisfies condition (U1) and, by the following lemma, condition (U3).

Lemma 6.6.3. *There are no mutually distinct sets A_1, A_2, A_3, and A_4 such that*

$$A_1 \cap A_2 = A_2 \cap A_3 = A_3 \cap A_4 \neq A_2 \cap A_4 = A_4 \cap A_1 = A_1 \cap A_3. \tag{6.2}$$

Proof. Suppose there exist pairwise distinct sets A_i, $i \in \{1, \ldots, 4\}$ that satisfy relation 6.2. Then we obtain that:

$$A_1 \cap A_2 \cap A_3 \cap A_4 = A_3 \cap A_4 = A_2 \cap A_4,$$

a contradiction. □

Thus D_α is a symbolic ultrametric on X if and only if D_α satisfies condition (U2). For example, if α is the accumulation map induced by a rooted phylogenetic tree then D_α is a symbolic ultrametric on X.

From the above considerations, it follows that Theorem 6.6.2 can also be proved using the main theorem of [10] (see also [48]) on the representation of symbolic ultrametrics by discriminating maps on rooted phylogenetic trees (together with Theorem 6.5.1).

We end this section with some remarks regarding further investigation on the model of accumulation phylogenies. It would be interesting to investigate variations on the model described - either by weakening one (or both) of the properties (**P1**) or (**P2**), or by allowing elements of S to be 'lost'. We describe briefly this last variation. Rather than requiring, for each arc $a = (u, v) \in A$ that

$$\overline{\alpha}(v) = \overline{\alpha}(u) \cup f(a)$$

we may weaken this to require merely that

$$\overline{\alpha}(v) = B \cup f(a)$$

where $B \subseteq \overline{\alpha}(u)$ to thereby model the situation where elements of S become lost over time. In this model we maintain (**P1**) and (**P2**) but allow some flexibility in the definition of $\overline{\alpha}$. As might be expected, one can say much less about the underlying hybrid phylogeny (V, A) from the induced accumulation map $\alpha = \overline{\alpha}|X$. However α does still convey some phylogenetic information - for example, suppose we know that $\mathcal{D} = (V, A)$ is a tree. Then if, for some subset Y of $X - \{x\}$, we have

$$\alpha(x) \subseteq \cup_{y \in Y} \alpha(y),$$

then this places a constraint on $\mathcal{D}$ - namely, the most recent common ancestor of the species in Y must also be an ancestor of x.

Bibliography

[1] L. Addario-Berry, M. Hallett, J. Lagergren, Towards Identifying Lateral Gene Transfer Events, *Pacific Symposium on Biocomputing (PSB'03)* **8**, 279–290, 2003.

[2] B.L. Allen and M. Steel, Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees, *Annals of Combinatorics* **5**, 1–13, 2001.

[3] J. Alroy, Continuous track analysis : a new phylogenetic and biogeographic method, *Systematic Biology* **44** (2), 153–178, 1995

[4] H.-J. Bandelt and A.W.M. Dress, Weak hierarchies associated with similarity measures – an additive clustering technique, *Bulletin of Mathematical Biology* **51**, 133–166, 1989.

[5] J. Bang-Jensen and G. Gutin, *Digraphs: Theory, Algorithms and Applications*, Springer-Verlag, London, 2001.

[6] M. Baroni, S. Grünewald, K. Huber, V. Moulton, Notes on hybrid phylogenies, in progress.

[7] M. Baroni, S. Grünewald, V. Moulton, C. Semple, Bounding the number of hybridisation events for a consistent evolutionary history, submitted. Departmental report UCDMS 2004/19, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, 2004.

[8] M. Baroni, C. Semple, and M. A. Steel (2003), A framework for representing reticulate evolution, *Annals of Combinatorics*, in press.

[9] M. Baroni and M. A. Steel, Accumulation phylogenies, submitted. Departmental report UCDMS 2004/1, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, 2004.

[10] S. Böcker and A.W.M. Dress, Recovering symbolically dated, rooted trees from symbolic ultrametrics, *Advances in Mathematics*, **138**, 105–125, 1998.

[11] M. Bordewich and C.Semple (2004), On the computational complexity of the rooted subtree prune and regraft distance, *Annals of Combinatorics*, in press.

[12] M. Bordewich and C. Semple – private communication, 2004.

[13] K. Bremer and H-E. Wanntorp, Hierarchy and reticulation in systematics, *Systematic Zoology* **28**(4) 624–627, 1979.

[14] D. Bryant and V. Moulton, NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks, *Molecular Biology and Evolution* **21**(2), 255–265, 2004.

[15] P.J. Cameron, *Combinatorics: topics, techniques, algorithms*, Cambridge University Press, 1996.

[16] E. Diday and P. Bertrand, An extension of hierarchical clustering: the pyramidal presentation, in: E.S. Gelsema and L.N. Kanal, Eds., *Pattern Recognition in Practice II*, North-Holland, Amsterdam, 411–423, 1986.

[17] W.F. Doolittle, Phylogenetic Classification and the Universal Tree, *Science* **284**, 2124–2128, 1999.

[18] A.W.M. Dress, D. Huson, and V. Moulton Analysing and visualizing sequence and distance data using SPLITSTREE, *Discrete Applied Mathematics* **71**, 95–109, 1996.

[19] J. Felsenstein, *Inferring phylogenies*, Sinauer Press, 2004.

[20] V.A. Funk, Phylogenetic patterns and hybridization, *Journal of the Missouri Botanical Garden* **72**, 681–715, 1985.

[21] D. Gusfield, S. Eddhu, C. Langley, Efficient Reconstruction of Phylogenetic Networks with Constrained Recombination, *Proceedings of the 2003 IEEE Bioinformatics Conference (CSB2003)*, Stanford, California, 363–374, 2003.

[22] D. Gusfield, S. Eddhu, C. Langley, The Fine Structure of Galls in Phylogenetic Networks, To Appear in *INFORMS J. of Computing* Special Issue on Computational Biology (2004).

[23] M.T. Hallett and J. Lagergren, New Algorithms for the Duplication-Loss Model, 4^th *Annual RECOMB'00*, Tokyo, Japan, 138–146, 2000.

[24] M.T. Hallett and J. Lagergren, Efficient Algorithms for Lateral Gene Transfer Problems, 5^th *Annual RECOMB'01*, Montreal, Canada, 149–156, 2001.

[25] M. Hallett, J. Lagergren, A. Tofigh, Simultaneous identification of duplications and lateral transfers, *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB'04)*, 347–356, 2004.

[26] F. Harrary, *Graph Theory*, Addison-Wesley, 1969.

[27] J. Hein, T. Jiang, L. Wang, K. Zhang, On the complexity of comparing evolutionary trees, *Discrete Applied Mathematics* **71**, 153-169, 1996.

[28] B. Holland and V. Moulton, Consensus Networks: A Method for Visualising Incompatibilities in Collections of Trees, in: G.Benson and R. Page, Eds., *Algorithms in Bioinformatics (WABI 2003)*, LNCS **2812**, 165–176 , 2003.

[29] J. Jansson and W.–K. Sung, Inferring a Level–1 Phylogenetic Network from a Dense Set of Rooted Triplets, *Proceedings of the Tenth International Computing and Combinatorics Conference (COCOON 2004)*, LNCS **3106**, Springer–Verlag, 2004.

[30] F.-J. Lapointe, How to Account for Reticulation Events in Phylogenetic Analysis: A Comparison of Distance-Based Methods, *Journal of Classification* **17**, 175–184, 2000.

[31] P. Legendre, Reticulate Evolution: From Bacteria to Philosopher, *Journal of Classification* **17**, 153–157, 2000.

[32] P. Legendre, Biological Applications of Reticulate Analysis, *Journal of Classification* **17**, 191–195, 2000.

[33] P. Legendre and V. Makarenkov, Reconstruction of Biogeographic and Evolutionary Networks Using Reticulograms, *Systematic Biology* **51**(2), 199–216, 2002.

[34] P.J. Lockhart, P.A. McLenachan, D. Havell, D. Glenny, D. Huson, and U. Jensen, Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition, *Ann. Missouri Bot. Gard.*, **88**(3) 458–477, 2001.

[35] W.P. Maddison, Molecular Approaches and the Growth of Phylogenetic Biology, in: J.D. Ferraris and S.R. Palumbi, Eds., *Molecular Zoology: Advances, Strategies, and Protocols*, Wiley–Liss, New York, 1996.

[36] W.P. Maddison, Gene trees in species trees, *Systematic Biology* **46**(3), 523–536, 1997.

[37] B. Mirkin, *Mathematical Classification and Clustering*, Kluwer Academic Publishers, Dordrecht, 1996.

[38] B.M.E. Moret, L. Nakhleh, T. Warnow, C. Randal Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme, Phylogenetic Networks: Modeling, Reconstructibility, and Accuracy, *IEEE Transactions on Computational Biology and Bioinformatics* **1** (1), 2004.

[39] L. Nakhleh, J. Sun, T. Warnow, C.R. Linder, B.M.E. Moret, and A. Tholse, Towards the Development of Computational Tools for Evaluating Phylogenetic Network Reconstruction Methods, *Proceedings of the Eighth Pacific Symposium on Biocomputing (PSB'03)*, 8:315-326, 2003.

[40] L. Nakhleh, T. Warnow, C. Randal Linder, Reconstructing Reticulate Evolution in Species – Theory and Practice, *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB'04)*, 337 - 346, 2004.

[41] G. Nelson, Reticulation in cladograms, in: N.I. Platnick and V.A. Funk, Eds., *Advances in Cladistics*, vol. 2, Columbia University Press, 105-111, 1983.

[42] New Zealand Plant Species Radiation Group page http://awcmee.massey.ac.nz/ NZ_Plant_Species_Radiation_Group/projects

[43] R.D.M. Page and M.A. Charleston, Trees within trees: phylogeny and historical associations, *Trends in Ecology and Evolution* **13** (9), 356–359, 1998.

[44] W.L. Perry, D.M. Lodge, and J.L. Feder, Importance of Hybridization Between Indigenous and Nonindigenous freshwater Species: An Overlook Threat to North American Biodiversity, *Systematic Biology* **51** (2), 255–275, 2002.

[45] L.H. Rieseberg, The role of hybridization in evolution: old wine in new skins, *American Journal of Botany* **82** (7), 944–953, 1995.

[46] D.F. Robinson and L.R. Foulds, Comparison of phylogenetic trees, *Mathematical Biosciences* **53**, 131–147, 1981.

[47] F.J. Rohlf, Phylogenetic Models and Reticulations, *Journal of Classification* **17**, 185–189, 2000.

[48] C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, 2003.

[49] P.E. Smouse, Reticulation inside the species boundary, *Journal of Classification* **17**, 165–173, 2000.

[50] P.H.A. Sneath, Reticulate Evolution in Bacteria and Other Organisms: How Can We Study It?, *Journal of Classification* **17** 159–163, 2000.

[51] Y.S. Song, On the Combinatorics of Rooted Binary Phylogenetic Trees, *Annals of Combinatorics* **7**, 365-379, 2003.

[52] Y.S. Song and J. Hein, Parsimonious reconstruction of sequence evolution and haplotype blocks: finding the minimum number of recombination events, in: G.Benson and R. Page, Eds., *Algorithms in Bioinformatics (WABI 2003)*, LNCS **2812**, 287–302, 2003.

[53] Y.S. Song and J. Hein, On the Minimum Number of Recombination Events in the Evolutionary History of DNA Sequences, *Journal of Mathematical Biology* **48**(2), 160–186, 2004.

[54] B. Snel, P. Bork, and M.A. Huynen, Genome phylogeny based on gene content, *Nature Genetics* **21**, 108-110, 1999.

[55] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis, Phylogenetic Inference, in: D.M. Hillis, C. Moritz, B.K.Mable, Eds., *Molecular Systematics*, Second Edition, Sinauer Associates, Sunderland, Massachusetts, 407–514, 1996.

[56] L. Wang, K. Zhang, and L.Zhang, Perfect phylogenetic networks with recombination, *Journal of Computational Biology* **8**, 69–78, 2001.

[57] H.-E. Wanntorp, Reticulated cladograms and the identification of hybrid taxa, in: N.I. Platnick and V.A. Funk, Eds., *Advances in Cladistics*, vol. 2, Columbia University Press, 81–88, 1983.

List of Symbols

$<_{\mathcal{D}}$	strict partial order on the vertices of the digraph $\mathcal{D}$, 8
$\leq_{\mathcal{D}}$	partial order on the vertices of the digraph $\mathcal{D}$, 8
α	the accumulation map on X , 112
$\overline{\alpha}$	the accumulation map on V , 112
$B(n)$	the collection of binary phylogenetic trees with n leaves, 11
$c(v)$	the cluster corresponding to v, 22
$c(\mathcal{H})$	the set of clusters of $\mathcal{H}$, 22
$d^-(v)$	the in–degree of v, 7
$d^+(v)$	the out–degree of v, 7
$\mathrm{diam}_{rSPR}RB(n)$	the diameter of the metric space $(RB(n), d_{rSPR})$, 14
$d_{rSPR}(\mathcal{T}, \mathcal{T}')$	the rSPR distance between $\mathcal{T}$ and $\mathcal{T}'$, 11
$h(\mathcal{H})$	the hybridization number of $\mathcal{H}$, 20
$h(\mathcal{P})$	the hybridization number of a minimal hybrid that displays a collection $\mathcal{P}$ of trees, 59
$h_r(\mathcal{P})$	the hybridization number of a minimal regular hybrid that displays a collection $\mathcal{P}$ of trees, 59
$h_r^+(\mathcal{P})$	the hybridization number of a minimal regular hybrid that displays a collection $\mathcal{P}$ of trees and has extra-leaves, 59
$h(\mathcal{T}, \mathcal{T}')$	the hybridization number of a minimal hybrid that displays the trees $\mathcal{T}$ and $\mathcal{T}'$, 59
$h_r(\mathcal{T}, \mathcal{T}')$	the hybridization number of a minimal regular hybrid that displays the trees $\mathcal{T}$ and $\mathcal{T}'$, 59
$h_r^+(\mathcal{T}, \mathcal{T}')$	the hybridization number of a minimal regular hybrid that displays the trees $\mathcal{T}$ and $\mathcal{T}'$ and has extra-leaves, 59

$H(\mathcal{C})$	the cover hybrid of $\mathcal{C}$, 22
$\mathcal{H}[\mathcal{T}_1, \mathcal{T}_2]$	the cluster union hybrid of $\mathcal{T}_1$ and $\mathcal{T}_2$, 41
$\mathcal{L}(\mathcal{H})$	the leaf set of $\mathcal{H}$, 20
$m(\mathcal{T}, \mathcal{T}')$	the cardinality of a maximum agreement forest -1, 71
$m_g(\mathcal{T}, \mathcal{T}')$	the cardinality of a maximum good agreement forest -1, 72
$RB(n)$	the collection of rooted binary phylogenetic trees with n leaves, 14
rSPR	rooted subtree prune and regraft operation, 15
$Reg(X)$	the collection of regular hybrids on X, 25
X_{triv}	the set of trivial clusters of X, 22

Index